Revolving Drivers: Data Mining and Discovering the Causes of Driver Turnover

by

Adam G.M. Pazdor

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada

December 2019

Thesis advisor

Dr. Carson K. Leung

Author

Adam G.M. Pazdor

Revolving Drivers: Data Mining and Discovering the Causes of Driver Turnover

Abstract

Turnover (or Churn) is a great concern to every industry. Employees who leave represent hours of training wasted and the expense of hiring a replacement, something undesirable for any business. Few industries experience the problem as acutely as the trucking industry, where turnover rates have been as high as 90%. Uncovering the underlying reasons that are behind why so many drivers leave their jobs is a point of priority for many trucking companies. A solution, or even an explanation, could mean hundreds of training hours and thousands of dollars saved. In this M.Sc. thesis, I examine real-life data from a trucking company and use a random forest model to understand the driver turnover situation.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I would like to thank Dr. Carson K. Leung for his support as my supervisor during my time as a graduate student. It was because of him that I chose to pursue a Master's degree at all, and he has shown infinite patience throughout my attempts to get a thesis completed. He has also been a great supporter of my fledgling teaching career, and provided me the opportunity to work on the industrial project that ultimately became this thesis.

I would also like to thank my examining committee members, Dr. Yang Wang (Computer Science) and Dr. Liqun Wang (Statistics) for their time in reviewing and analyzing my thesis. I thank the University of Manitoba and NSERC for making my industrial project possible.

I would like to thank the students in the Data Science, Database, & Data Mining lab for always creating a friendly and welcoming environment, especially ex-lab member Dr. Fan (Terry) Jiang, who was an inspiration to me and in whose footsteps I am attempting to follow.

I thank my partner on the project, Joglas Souza, as well as Grant Barkman and Wing Kwong at DecisionWorks. I also thank Dave F., Adrian F., and Robert M. for their expertise as workers in the trucking industry.

Finally, I would like to thank my parents for supporting me through this degree, especially when it took a lot longer than any of us expected.

Adam G.M. Pazdor
B.Sc.(Maj.), The University of Manitoba, 2014

The University of Manitoba

December 06, 2019

Chapter 1

Introduction

The problem of employee turnover (employees voluntarily terminating their positions) has puzzled both employers and researchers alike for over a hundred years. In their literature review on the subject in 2017, Hom et al. [HLSH17] found articles dating back to the early 1900s, including one from 1906 on the turnover problem in the postal service. Since then, there have been hundreds of papers approaching the problem from many different angles, all in search of a solution that seems particularly elusive. While published research only goes back about a century, it is reasonable to speculate, given the issue's long documented history, that turnover has been a problem in the workforce (across all sectors) for as long as there has been a workforce.

Average turnover rates vary wildly across different industries and different time periods, making accurate comparison data hard to come by. Numbers from the U.S. Bureau of Labour Statistics [US 19] suggest that many industries have a voluntary turnover rate between 10 and 20%, though for some industries, that number can get much higher. Nursing, for instance, has experienced rates north of 80% [Wel18], and

hospitality is reportedly even worse. For the transport industry, the focus of this thesis, the turnover rate was nearly 100% in 2018 [McN18] (meaning as many people quit as were hired in the same period), making it an industry where turnover is a serious problem.

For an answer to this difficulty, we turn to the fields of data mining and machine learning. Data mining, the process of discovering implicit, previously unknown, and potentially useful patterns in data, is ideally suited to looking at turnover data on a large scale and uncovering the patterns that separate employees who leave from those who do not. We use machine mearning to train a computer to automatically process employee data and make predictions about unseen employees. While it certainly of interest for academics to uncover the reasons behind employee turnover, it is more valuable to the industry if we can provide a method by which they can assess an individual driver's risk of turnover.

Much of the existing work on this topic focuses on the former; seeking simply to identify the factor or factors which contribute most to the problem. We contend that even knowing the most important factor or factors is not sufficient. Turnover is evidently a complex issue, and blanket solutions applied to all employees ignore the intricacies by which the factors of turnover interact. While it is impossible (or at the very least impractical) to create a model individually tailored to every employee, by considering multiple factors in conjunction we hope to create a model that can provide predictive power to employers when they apply their own employees to it.

Any cursory glance at this subject poses the following three immediate questions:

1. What exactly *is* employee turnover (i.e. how is it measured)?

This question appears straightforward, but has nuance of its own. It is discussed in detail in Section 2.1, but in brief, turnover is a measure of the rate at which employees leave the company for which they work.

2. Why is turnover such a big problem?

   This question has many answers. The economic reason is perhaps the most obvious. When employees leave, employers must spend additional time and resources locating, hiring, and training a replacement. This is particularly damaging when employees leave soon after their hire date. A company's investment in an employee is largely up-front (apart from perpetual items such as salary and other benefits), whereas the benefit the employee provides to their employer is usually more long-term. A person who leaves a company before those training costs have been "recouped" in work done is an extra expense. According to Allen et al. [ABV10], hiring and training costs range from 90% to as high as 200% of an employee's annual salary. This means that an employee who leaves after just one year could have potentially cost the company three years' worth of expenses, all of which is now wasted.

   Turnover contagion [FMH+09] is another reason that this issue is as serious as it is. Employees who witness other workers quitting in droves become more likely to quit themselves, compounding the problem. Worse still, when employees leave a company, their skill set makes them likely candidates to be hired by a competitor, which only does further damage to their original employer [AGZ09].

   Frequent turnover also means that companies end up with more and more drivers who have little experience. For a recent historical example, consider

the horrific crash on April 6th, 2018, which killed sixteen members of the Humboldt Broncos when their bus was crashed into by a semi. The driver who caused the crash had only two weeks of training prior to the accident and had only held a license for one year [MH18]. Accidents of this magnitude are rare, but may become more frequent if companies cannot reliably put experienced drivers on potentially unsafe roads.

3. What causes employee turnover?

   Answering this question has been the subject of dozens of research papers over more than a century, and has no definitive answer. It is the goal of this project to uncover the root causes of turnover in the data we have, and then hopefully to be able to extrapolate that to other companies and industries.

   As a preview, our model addresses the rampant single-factor issue, examining more than a dozen factors in concert. We identify the most important variables across many different iterations of the model, and yield a predictive algorithm that performs significantly better than chance under most conditions.

1.1 Thesis Statement

Driver turnover is a long-standing problem, and one for which there has never been a definitive explanation, let alone a solution. The fact that there are still studies being done on this issue after more than a century of research is testament to this. A new approach is needed to discover the root of the problem on even the smallest scale.

For this MSc research, the goal is to develop a model that can predict the

likelihood of a driver to turnover with a reasonable degree of accuracy. My key contribution is the unique combination of feature variables, model hyperparameters, and data views used, which resulted in a model that can be successfully applied to an entirely different class of driver than the one on which it was trained.

Questions to be answered in this thesis:

1. What combination of feature variables, hyperparameters, and data manipulation results in the most accurate driver prediction model?
2. Can a model trained on one type of driver successfully predict the behaviours of a different type of driver?

1.2 Thesis Organization

The remainder of the thesis is organized as follows. More background information on turnover and the metrics used is provided in Chapter 2. Chapter 3 contains a review of related works on turnover, both in general and in the trucking industry in particular. We also summarize how this thesis research compares to existing work.

In Chapter 4, the creation of the Random Forest model is described. We go into detail about the datasets that were used to train, test, and validate the model. We also explain the features used in the model, and how the model's hyperparameters were tuned for optimal performance.

Chapter 5 contains the experimental results of multiple tests of the model, including our best-case scenario, as well as several other configurations that were of interest.

This chapter also contains a discussion on the broad conclusions that can be drawn from the results of the model.

Finally, in Chapter 6, we have the research conclusions and a discussion of future work that could be done to extend this project on driver turnover.

Chapter 2

Background

2.1 Turnover

Employee turnover refers to the rate at which employees leave their company. It is usually expressed as a percentage of employees who left over some time period (typically one year). One common calculation [Pav16] is:

$$TurnoverRate = \frac{EmployeesWhoLeft}{(StartingEmployees + EndingEmployees)/2} \times 100\% \qquad (2.1)$$

So, for example, a company that starts a year with 100 employees, ends it with 150, and had 25 employees leave during that time, would have a turnover rate of $\frac{25}{(100+150)/2} = 1/5 = 20\%$.

It is important to note that the term "employee turnover" is often used interchangeably with "voluntary employee turnover", which is a measure specifically of the rate at which employees choose to leave their jobs (as opposed to being terminated by the employer or some other external reason). For the remainder of this

thesis, the terms "employee turnover" and "turnover" will be used to refer to voluntary employee turnover.

2.2 Data Mining

Data mining is the process of extracting from data information which is useful, implicit, and previously unknown [FPS+96]. It is used increasingly in every field, particularly in areas where there are large volumes of data which are hard for humans to analyze directly. In our research, we use data mining techniques to help us analyze the transport data we have, in hopes of finding patterns that can tell us the root causes of driver turnover.

2.3 Machine Learning

Machine learning (ML) is a field related to data mining, referring to the study of algorithms and models where a computer is "trained" to do analysis without explicit human instruction. Machine learning models vary wildly, but almost all depend on a set of training data for the model to practice on, and then a test of new testing data for the model to use to verify its predictive power.

Where data mining is more concerned with discovering existing patterns, the purpose of machine learning is for prediction. We use a combination technique in our research; using data mining to discover the key factors in driver turnover and then using a machine learning algorithm to predict turnover in new drivers.

2.3.1 Supervised Learning

The term Supervised Learning refers to a specific type of machine learning model. Namely, it refers to data which has already been pre-classified before being given to the model. The model then takes all of the inputs and the desired output and constructs a function which predicts that output as closely as possible. The name puts it in contrast with Unsupervised Learning, which is a type of ML model where the model is given only inputs with no desired output for each.

We used a Supervised Learning model in our research because the output (whether or not the driver left the company) was known to us in advance and thus we could include it in the training data.

2.4 Decision Tree

A Decision Tree is a structure similar to a flowchart used in any form of decision making or decision analysis, and is a commonly-used tool in data mining as a classification algorithm and machine learning as a supervised algorithm. Usually drawn left-to-right or top-to-bottom, a Decision Tree is made up of multiple nodes, representing decisions, and lines, representing the paths made by those decisions. Sink nodes (leaves) represent the final state or classification of the end of a path.

An extremely simple Decision Tree is shown in Figure 2.1. In it, we follow a nurse's decisions as they try to determine if a patient has a cold or the flu. Our algorithm uses Decision Trees in a similar way, putting the most important attribute first, and then working down to a decision at the end. As with our illustrative example, while

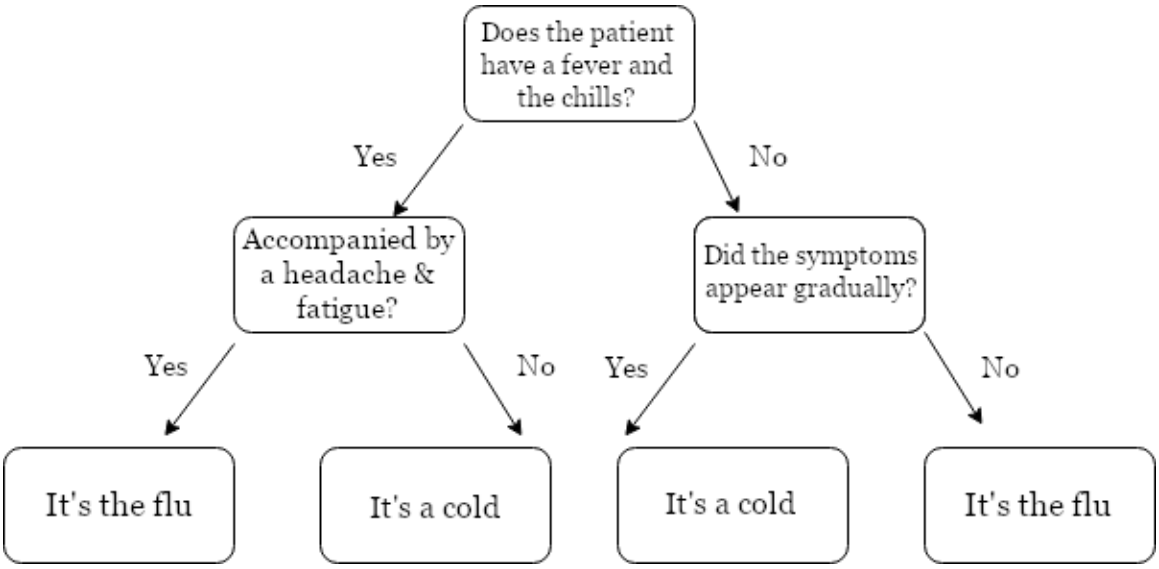

Figure 2.1: Sample Decision Tree

there are multiple leaf nodes, there are ultimately only two outcomes (a driver staying or leaving).

2.5 Random Forest

The *Random Forest* model is a data mining and machine learning model for data classification [Bre01]. It is based on the creation of multiple *Decision Trees* (hence the name), each of which uses some subset of the feature variables in the training dataset. The collection of decision trees is then collectively used to make an overall decision in the classification step. The analysis done in this thesis uses the *Random Forest* model.

*Random Forest*s are particularly appropriate for this style of problem, as we have a large number of feature variables, and we purposely make no up-front assumptions

about which ones are more important. The fact that a *Random Forest* takes multiple samples of the feature variables in its decision allows for a natural pruning of the variable list as the model is applied to the training data.

2.5.1 Partial Dependence Plots

One of the disadvantages to a Random Forest model is that it is a "black box". Once the model is created, it is hard to look inside and see how each individual feature variable impacted the final product. Partial dependence plots (PDPs) [Fri01] provide a little insight; they show the marginal effect of one or two features on the model.

As a preview, we use PDPs in Chapter 5 to visualize the impact of our most important feature variables. We built our PDPs using Python's PDPBox library [PDP19].

2.6 F1 Score

One of the primary pieces of the metric used to evaluate our model is the *F1-score* [Pow07] (also called an F-score or F-measure). An *F1-score* is one way of summarizing a binary classification table (aka Confusion Matrix), as seen in Table 2.1.

For our research, driver turnover is our binary variable; our positive class is "the

Table 2.1: Binary Classification

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

driver left the company" and our negative class is "the driver did not leave the company".

The True Positive value represents the number of correct predictions of the positive class. In our case, it is the number of drivers who were correctly predicted to quit.

The False Positive value represents the number of incorrect positive predictions. It is also known as Type I error. In our case, it is the number of drivers who were predicted to turnover but remained at the company.

The False Negative value represents the number of incorrect negative predictions. It is also known as Type II error. In our case, it is the number of drivers who were predicted to remain but left the company.

The True Negative value represents the number of correct predictions of the negative class. In our case, it is the number of drivers who were correctly predicted to stay.

The sum of True Positive and True Negative is the number of correct predictions, and the sum of False Positive and False Negative is the number of incorrect predictions.

Many evaluation metrics can be derived from this table, but for brevity, only *Precision*, *Accuracy*, and *Recall* are explained here.

2.6.1 Precision

Precision (also known as Positive Predictive Value) is the ratio of the number of values correctly predicted to be positive to the number of total positive predictions. In simple terms, it is the answer to the question "Of all the times you [the model]

predicted that X would fall into the positive category, how often were you right?". It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

In this thesis, it represents the fraction of drivers who left the company out of all those who were predicted to leave.

2.6.2 Recall

Recall (also known as Sensitivity) is the ratio of the number of values correctly predicted to be positive to the number of total positive values. In simple terms, it is the answer to the question "What percentage of the positive X values did the model predict correctly?". It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

In our thesis, it represents the fraction of drivers who where predicted to leave the company out of all those who left.

2.6.3 Accuracy

Accuracy is the ratio of the number of values correctly predicted to the total number of cases. It is perhaps the simplest summary statistic of the binary classification table. In simple terms, it is the answer to the question "What percentage of all the values did the model predict correctly?". It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

2.6.4 F1-Score

Finally, the F1-Score is a combination of Precision and Recall. There are multiple versions of the F1-score based on what relative weights are given to Precision and Recall in the calculation. The traditional F1-Score (the one used in this thesis), is the harmonic mean of Precision and Recall:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.5}$$

2.7 Receiver Operating Characteristic

Receiver Operating Characteristic Graphs (usually abbreviated as ROC Graphs) are a visual evaluation method based on the Confusion Matrix (Table 2.1) [Faw06].

2.7.1 ROC Space

An ROC graph is two-dimensional, where the True Positive Rate (TPR) is plotted on the Y-axis and the False Positive Rate (FPR) is plotted on the X-axis. An example of ROC space is shown in Figure 2.2, which includes a sample ROC curve (the green line).

Each point in ROC space represents an instance of a Confusion Matrix. The best possible classifier would have a TP rate of 1 and a FP rate of 0 (in other words it would occupy the top-left corner of the ROC space). A classifier guessing at random would have a point somewhere on the diagonal line.

When testing a model in ROC Space, we want to see points far away from the diagonal; while anything below the diagonal is technically a worse-than-random result,

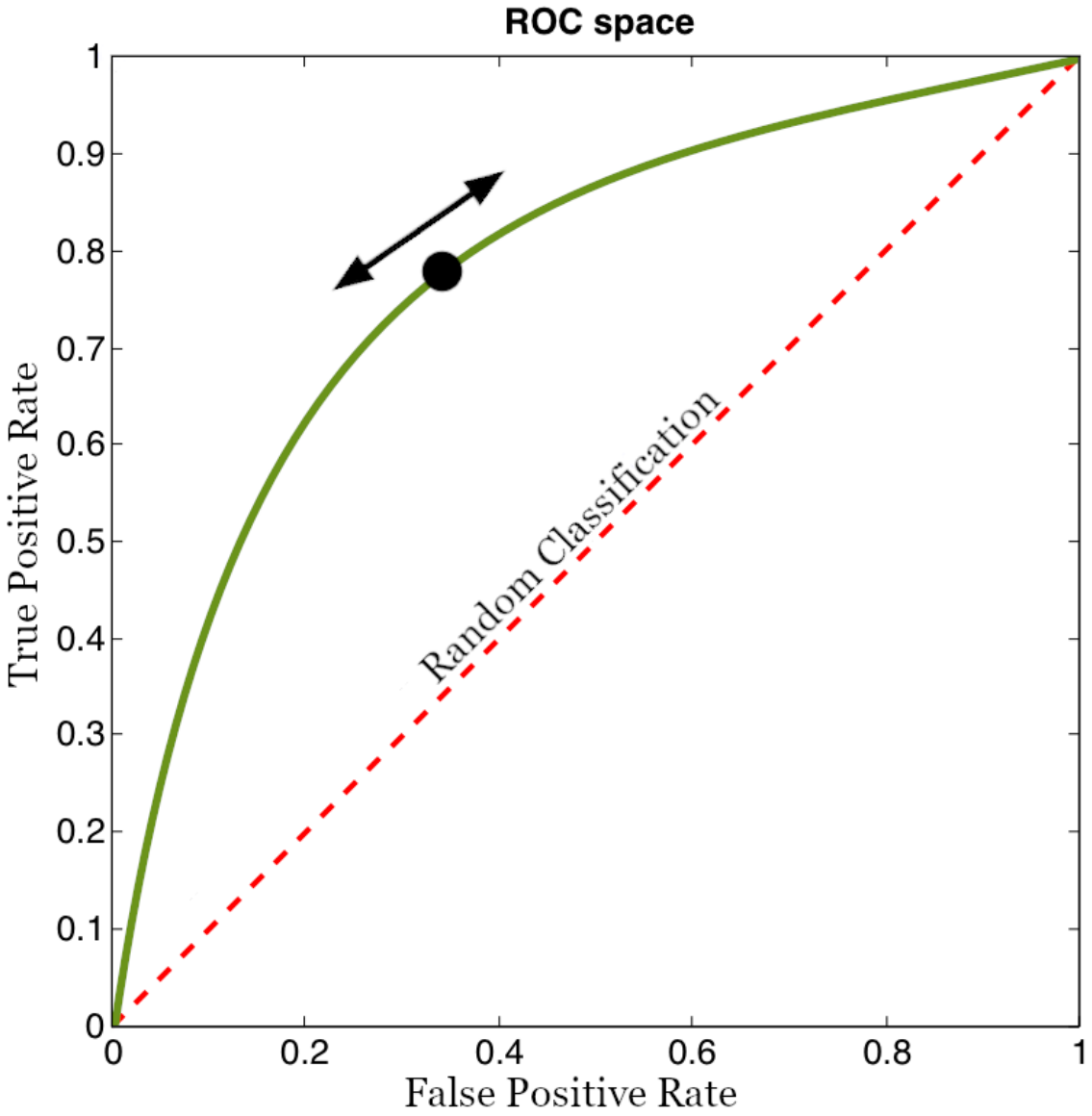

Figure 2.2: ROC Space

a model that tends toward the bottom-right corner can simply be inverted (flipping positive predictions for negative and vice versa) and it then becomes a better-than-random predictor.

2.7.2 ROC Curve

Many binary classifiers, including Random Forest, perform their classification by calculating a score - a probability that some sample value Y falls into the positive case - and then comparing that score to a threshold. The Random Forest we used, for example, has a threshold value of 50%, meaning that if the model gives some driver D a 50% probability of turning over, it predicts that driver to leave. Otherwise, it predicts the driver to stay.

An ROC curve is generated by taking a single model and repeatedly plotting its results with different values for the threshold. Over many iterations, this series of points produces an ROC curve. Mathematically speaking, the curve plots $TP(T)$ versus $FP(T)$:

$$TPR(T) = \int_T^\infty f(x)dx \tag{2.6}$$

$$FPR(T) = \int_T^\infty g(x)dx \tag{2.7}$$

where X is the continuous random variable (i.e. the score) created by the classifier, and thus $f(x)$ and $g(x)$ are the probability density functions for an instance being classified as positive or negative, respectively. A visualization of this is shown in Figure 2.3, where the vertical black bar in the upper-left graph is the threshold. Moving the black bar moves the black dot on the ROC Graph (see Figure 2.2), creating the curve.

The best possible ROC Curve, as seen in Figure 2.4, makes a right-angled triangle with the diagonal, representing the case where the PDF for TPR and FPR have no overlap (and our classifier is thus perfectly accurate when we get to a T value between the two curves).

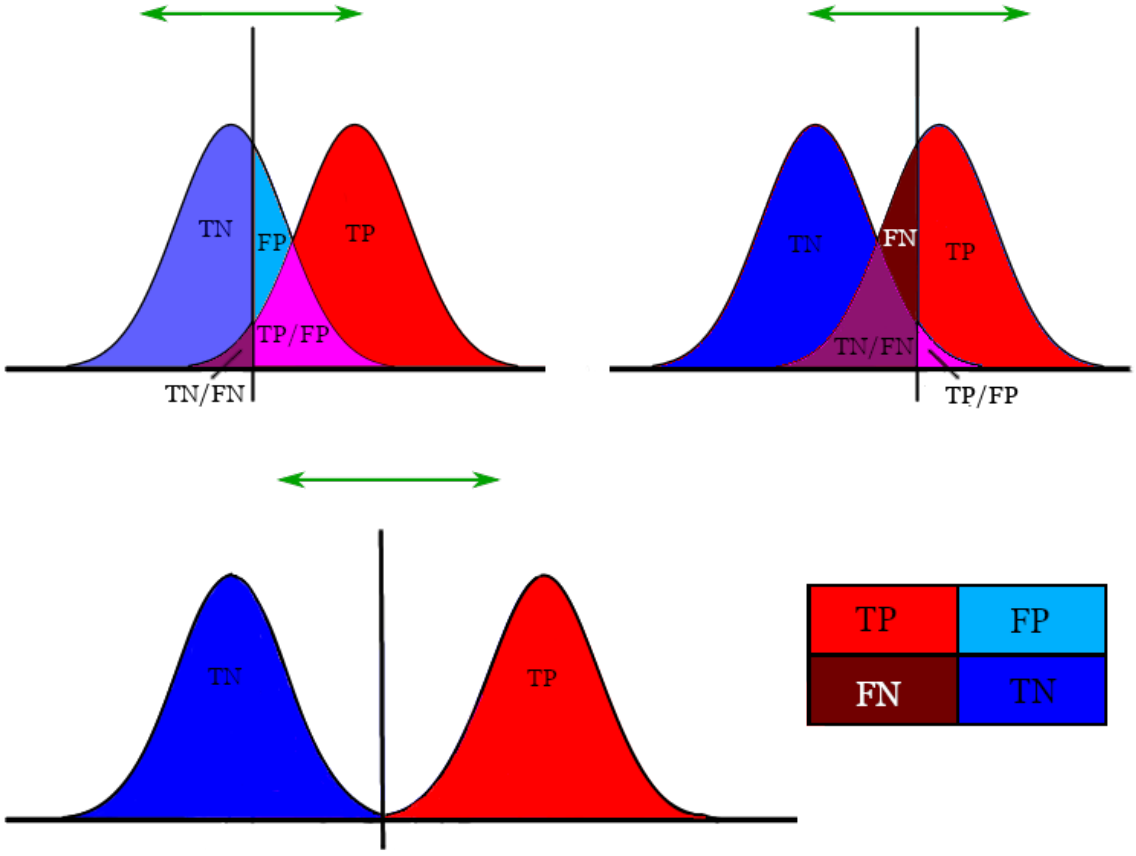

Figure 2.3: ROC Curve Explanation

In order to summarily evaluate an ROC curve, one metric is to calculate the area under the curve (AUC). The AUC is a useful metric, as it provides an answer to the question: "What is the probability that a randomly chosen positive instance $Z+$ will be scored higher (i.e., given a higher probability to be in the positive class) by the model than a randomly chosen negative instance $Z-$?" [Faw06]. In the perfect case, the probability of such a thing is 100%, as the model will always mark the positive cases as positive and the negative cases as negative (thus always ranking the positive cases higher). For a classifier operating at random, the AUC would be 50%, as any

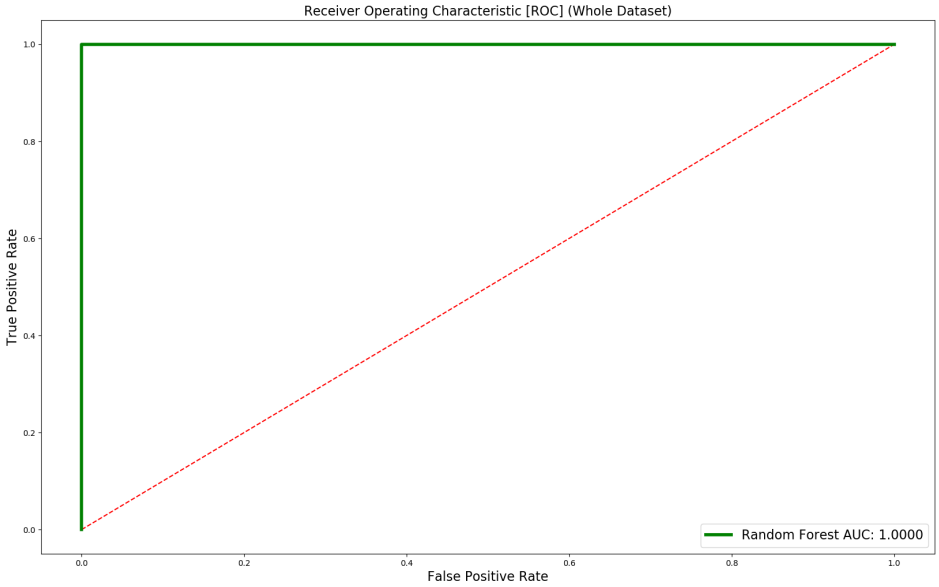

Figure 2.4: ROC Ideal Curve

one sample would have an even chance of being ranked above any other.

In this thesis, we also create ROC Curves and calculate the area under the ROC curve. To do so, we used Python's scikit library [PVG+11].

2.8 Summary

In this chapter, we outlined necessary background information for our research, including the definition of turnover, the Random Forest model, and two evaluation metrics: F1-score and ROC-AUC.

It is worth nothing that neither of the explored metrics are beyond criticism. Both of them, for instance, have been criticized because they treat precision and

recall with equal weight [LJR08, Pow15, HC18]. In statistical terms, they treat Type I and Type II errors equally. This approach is not always practically appropriate; a missed diagnosis (Type II error) of an easily-treated but fatal disease is much more detrimental (to physical health if not mental health) than a false positive diagnosis of the same.

It is our belief that this particular criticism is not valid in the case of driver turnover, as it is approximately equally damaging to incorrectly predict a driver to stay (and thus not devote any resources towards keeping them) as it is to incorrectly predict a driver to leave (and thus waste resources trying to keep them when they were already staying). Our own evaluation metric, discussed below, combines ROC-AUC and F1-score, and thus shares this weakness, but should otherwise be more stable than either metric individually.

Chapter 3

Related Works

3.1 Employee Turnover

With turnover being a problem for virtually every industry (e.g. Hospitality, Retail, Nursing, Banking, Insurance), there has been no shortage of research done on the subject already. There have been more than 250 articles on turnover [HLSH17], the plurality of which have been published in the *Journal of Applied Psychology* [AHVM14].

While articles and studies exist in the early 1900s, the first glut of research took place during the 1960s and 1970s [HLSH17], where the most popular predictive model used Weighted Application Blanks (WAB) ([Bue64, Sch67, Cas76, FFL76]), though there were those who questioned those studies for being rarely cross-validated [SO74]. WABs focused on an employee's personal history rather than their relationship with the company and have largely fallen out of favour.

Hulin [Hul66] introduced features that would shape the way this research was

done for decades to come, including job satisfaction measures (e.g., [SKH69]), isolating voluntary turnover rather than including involuntary dismissal, and focusing on individual relationships rather than the aggregate (e.g., [BC55]). Hulin is also often credited with creating (or at least setting the ground for) the "standard research design" [Ste02]; a policy of collecting predictors at one point in time and turnover data at a later point.

We incorporated these principles into our research; our model ignores drivers who were terminated as opposed to leaving of their own volition, and we include analysis of drivers on an individual level in the results.

In what has been called the "most influential single paper on turnover" [HLSH17], Mobley outlined a process, which he theorized outlined how job dissatisfaction results in turnover [Mob77, MGHM79]. His key step involved the employee calculating the subjective expected utility (SEU) of leaving versus staying (including the costs associated with quitting and finding a new job). Mobley argued that an employee's decision to leave or stay may be based on the future as much as the present. An employee may remain in a job with which they are dissatisfied if they expect future benefits (e.g. a promotion), and conversely an employee may leave a good job if they feel they can get higher utility from other employment.

Mobley's model dominated the field for nearly two decades until Lee and Mitchell [LM94] proposed an alternative: the Unfolding Model. In their model, an employee leaving for better prospects is but one of four different "paths", and that people often quit after experiencing some sort of negative "shock", such as the birth of a child or being asked to do something unethical by their boss. The Unfolding Model has since

become the dominant method of explaining turnover in the psychology field [Hom11].

In contrast, the nature of the data we received to work with precluded directly implementing either Mobley's approach or the Unfolding Model, but they are included here for completeness and as a grounding for some of our proposed future work.

In more recent works, Sota et al. [SOK19] tested a neural-network-based model on the restaurant industry in Japan to some success, furthering the idea that the solution to this problem lies in complex models rather than simple solutions. Recent studies in China have also looked at the impact of employees' social media usage [LDP19, ZMXX19] on their turnover rate, suggesting that employees who engage with the company on social media are less likely to leave.

3.2 Turnover in Trucking Industry

As a subset of the large body of research done on turnover in general, this thesis is hardly the first to study this problem as it pertains to the trucking industry. According to the American Trucking Association, turnover (especially for large fleets) has been a problem for decades, and reached 94% in the first quarter of 2018 [McN18]. It improved somewhat in the first part of 2019, dropping to 87% for large carriers (smaller carrier held steady around 70%) [All19]. The reason cited for this is an increase in pay to truck drivers. While it seems obvious and intuitive that offering drivers additional pay incentivizes them to stay, many in the research field have looked for other explanations. Even existing empirical evidence disputes this simple solution; an exit poll in the 90s found that just 73 of 2,050 drivers cited pay as a reason for leaving [LT89].

In 2011, a study was done to try and identify the root causes of intentions to quit, focusing on the label of "job satisfaction", a combination of many factors about a driver's workplace not simply their rate of pay [ABS11]. In that paper, the authors cite an earlier work that says quitting has a significant financial impact on the driver, not just the company. This leads into their central focus; existing work on this topic largely focuses on the company side of things, often getting perspectives of managers rather than the drivers themselves [RLTT94].

The authors surveyed almost 500 drivers working for a single American trucking company. They measured several esoteric factors by asking the drivers about their job experiences, including recognition, fairness, and commitment. Their conclusion was that it is a complex issue with no simple answer, though driver perceptions about how fair their workplace is and how recognized they feel plays an impactful role.

As all the data we have for this project is objective rather than the drivers' subjective viewpoints, we could not implement these findings into our model, but we bear them in mind as a piece of the puzzle that may yet be missing.

In 2017, a study was done approaching this problem from a similar angle; namely driver confidence [HB17]. In it, Hoffman and Burks postulate that overconfidence may be a factor in turnover, as drivers who expect to be more productive are less likely to leave. Their research showed that overconfidence is rampant, with drivers on average predicting they will drive nearly 25% more miles than they actually do. However, their model found that disillusioning drivers would be worse for companies (though slightly better for the welfare of the drivers themselves). One other interesting result was from an exit survey they did of drivers who left the company during the study

period. They found that nearly half of responders were leaving the trucking industry altogether, and only 12% were going to another long-haul trucking job (Hoffman and Burks' research had been specifically on long-haul drivers) [HB17].

Most recently, in 2018, Conroy et al. [CRDG18] revisited the issue of payment, but from a new perspective. Namely, they did not consider how much a driver is paid relative to the industry or their peers, but rather relative to themselves. Conroy et al. called this "pay variability" and found that an increased amount of variance in a driver's pay over time has an impact on a driver's likelihood to turnover [CRDG18]. Variability is a particular issue in the trucking industry, as a driver's pay rate can fluctuate wildly depending on the number of miles they drive and the amount of hours they work. We use pay variability as one of our feature variables.

3.3 Existing Turnover Factors

Over the last century, driver turnover has been attributed to many causes, and even more combinations of those causes. WABs were popular when it was thought that turnover could be blamed solely on the employee [Sch67], but most modern research focuses more on what an employer can do to affect turnover rate.

Factors which have been thought to contribute to employee turnover include: Workplace conditions [Hul66] (or even just the perception thereof [KN73]), job satisfaction [SKH69], and perception of management [FH62]. Hellriegel and White [HW73] found "leavers" (employees who turnover) expressed dissatisfaction with their workplace environment: their shifts, performance reviews, pay, and how their talents were being used. This research was echoed by Massoni et all [MGS+19], who surveyed

Brazilian software developers who left their jobs on their job characteristics and satisfaction. In media, is it common to see pay being highlighted as the single most important factor [All19], but research suggests it is much more complicated than that.

Porter and Steers [PS73] suggested that it is not poor conditions by themselves that cause turnover, but rather the gulf between an employee's expectations and the reality of the job. Porter also suggested in a later work [DKP81] that it is worth putting weight on what sorts of employees are leaving (previous research had almost entirely focused purely on turnover *rate*), arguing that turnover is only (or at least most) damaging when it is highly talented employees who leave.

3.4 Summary

In this chapter, we outlined some of the large body of existing work that has been done on the problem of employee turnover, both in general and in the trucking industry in particular. The consensus among the research seems to be that there is no simple solution to turnover and the common belief that it is simply a case of drivers being underpaid does not hold up to scrutiny.

The conclusions drawn by those who came before us informed the way we created our model, including which feature variables we used. That said, many pieces of existing work focus on one or two predetermined supposed causes of turnover (or an aggregate item in the case of [CRDG18]), and sought to prove or disprove that it was a central reason for driver turnover. In contrast, our research looks at a wide variety of possible factors at once and narrows in on the few we experimentally show to be

more important.

Of note is the fact that, while there is certainly a massive body of work on the subject of turnover in general and driver turnover in particular, we did not find any directly comparable existing work to our own. To the best of our knowledge, there is no research that examined a large number of variables for hundreds of drivers in an attempt to create a predictive machine learning model. The works done on the subject in a post-Machine-Learning world either focused on demonstrating that a single variable was significant ([HB17, CRDG18]) or approached the problem from a different enough angle that the work is not directly comparable ([ABS11]).

Chapter 4

Model Creation

4.1 Overview

The research done for this thesis uses data from a single real-world trucking company, but the process we used can be applied to data from any trucking company (or indeed most companies in general) with minor edits. We provide a high-level overview of our Model Creation steps here.

1. Clean the data

   The specific ways this is done depend on the data, but will likely include some combination of the following:

   (a) Removing null/invalid rows/columns from the data

   (b) Removing employees who were terminated rather than quit

   (c) Divide up the data according to employee type, if desired (in our example, we separate the data for Owner-Operator drivers and Company drivers)

2. Develop feature variables

   Some of these will come naturally from the data (e.g. age) whereas others will need to be calculated (e.g. average weekly payment). This step can be very esoteric and depend heavily on the nature of the data. We tested many initial variables and trimmed our list down when some of them were shown to not have significant impact on the model.

3. Choose and implement a classification model (e.g. Random Forest)

4. Optimize the model's hyperparameters using a subset of the data

5. Optimize the way the model is fed the data.

   This step is also highly dependent on the nature of the data, but in general it involves splitting up the data along lines which are suspected to be significant. For example, if the company has a three-month training period, it may be of interest to divide up the data according to employees who quit during that period versus those who left after it was over. In effect, modifications made to the data in this way become additional "feature variables".

6. Run the model multiple times on the data according to each data view (following the example above, this would run it once on all of the data, and then once on the data divided up by whether or not the employee left before the training period ended).

7. Calculate a score for each run of the model based on some desired evaluation metric.

4.2 Data Preparation

All of the data for this project was obtained courtesy of a real-life North American transport company, hereafter referred to as Company XYZ. For privacy and proprietary reasons, the realm name of the company cannot be disclosed. According to XYZ's internal numbers, their turnover rate fluctuates around 20%, which puts them on the low end for the transport industry, but still high overall.

Company XYZ provided us with data from as far back as 2013. Figure 4.1 shows an overview of all the driver data we received. The initial analysis of the data revealed some points of interest.

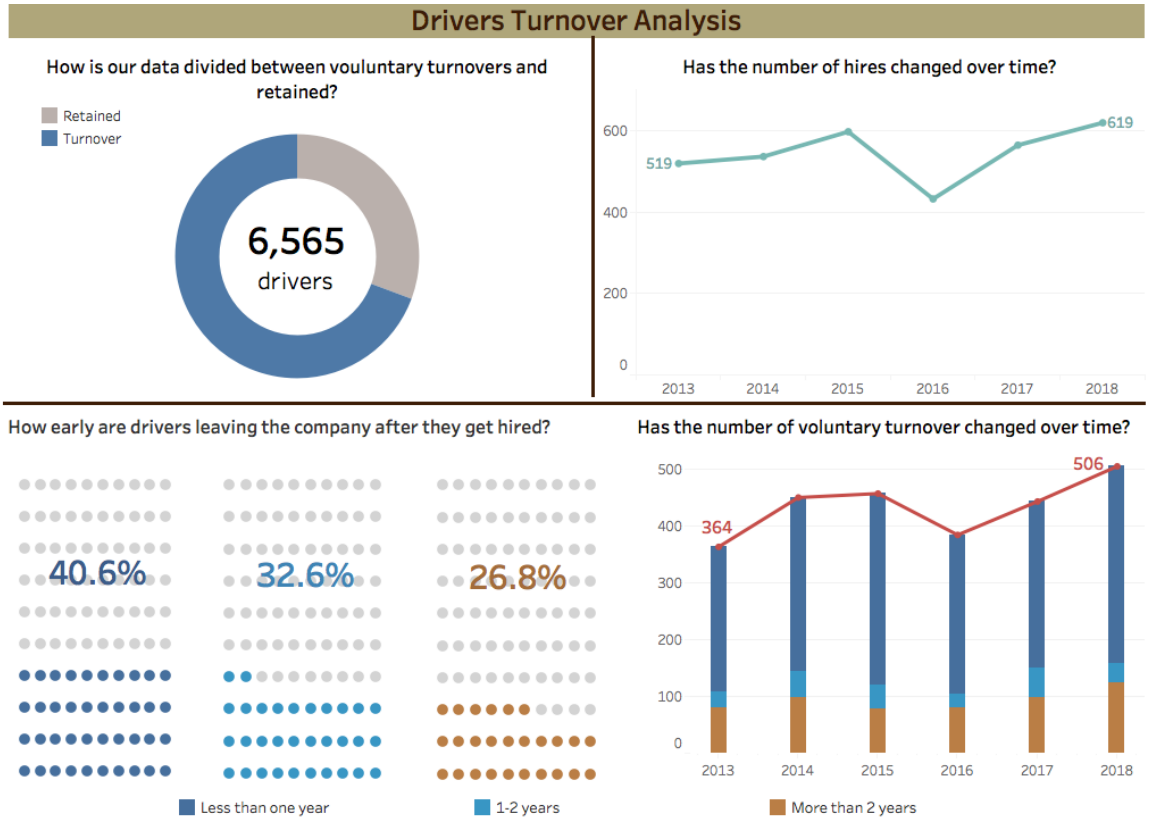

Figure 4.1: Initial Data View

1. Of the 6,565 drivers for whom we had data, nearly three-quarters had left the company. This suggested to us that we might need to use up-sampling or down-sampling to make the model work effectively (though this turned out to be unnecessary).

2. The turnover numbers seemed to be correlated with the hiring numbers. Turnover had an upward trend between 2013 and 2018, with the only dip being in 2016. This parallels the pattern of number of new hires, suggesting that 2016 was not special except in that fewer drivers were hired and thus there were fewer possible drivers to turnover.

3. We can see that the vast majority of drivers who leave quit within the first two years (nearly 75%), and a large percentage (40%) leave within one year of being hired. This data point illustrates the importance of solving this problem; the cost of training a new driver is mostly up front, meaning drivers who leave soon after training are more expensive to replace, relative to the amount of work they have done for Company XYZ.

4.2.1 Data Cleaning

The data we had for the 6,565 drivers was incomplete in some places. By the time we had decided on the shape of our model, it was clear that not all of the drivers would be usable. The following steps were undertaken to clean the data and create our final dataset.

- Null columns for all drivers were removed

- Columns where all drivers had the same entry were removed
- Rows containing inconsistent or unverifiable data (in collaboration with Company XYZ) were removed
- Entry-based datasets were aggregated into summary columns
- Drivers who were terminated rather than quit were removed
- As we did not have live data, drivers who were still with the company as of December 31, 2018 were considered to be retained

4.2.2 Active Dataset

Figure 4.2 shows the data for the 1,169 drivers actually used to train and test our model. This data contains drivers hired from 2015 onwards, and only contains Company (drivers who were employed directly by Company XYZ and drove trucks owned by XYZ) and Owner-Operator (drivers who owned their own truck, but still worked for XYZ) drivers, as they were the largest subsections of driver types.

The trends for our subset of data are mostly the same as in the larger dataset, with two notable exceptions.

1. The percentage of drivers who left within one year of hiring jumped from a plurality to a large majority (over three-quarters).
2. The ratio between turnover drivers and retained drivers is much more balanced, which led us to think that we would not need to up-sample or down-sample the datasets to get a viable model.

Table 4.1: Raw vs Cleaned Data

Dataset	Item	Raw	Cleaned
Drivers	Records	6,565	1,169
Driver	Attributes	196	93
Legs	Records	4.5 million	3.5 million
Legs	Attributes	42	33

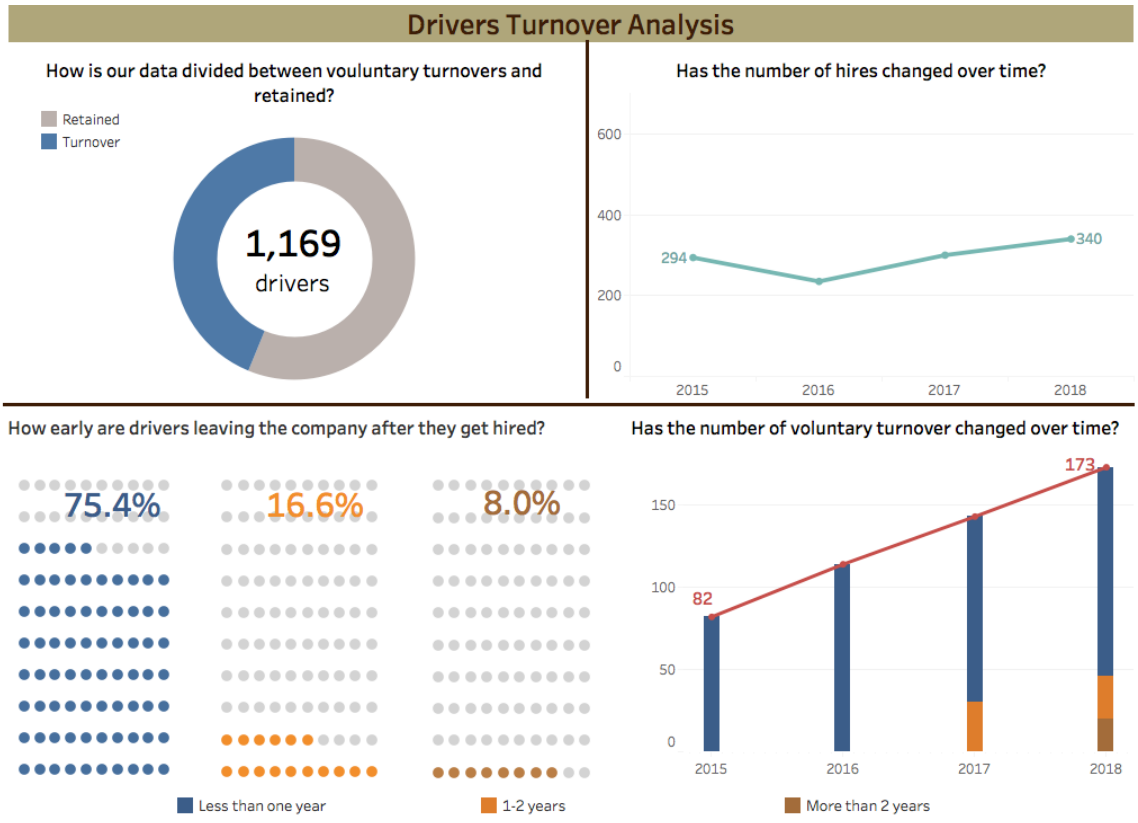

Figure 4.2: Final Data View

4.3 Feature Engineering

The data we received was in two broad categories: data about drivers and data about trips (aka legs) those drivers took for the company. The leg data was aggregated, and the aggregate values were used in our analysis.

Our final list of feature variables, in decreasing order of importance, consisted of the following:

- Years at company
- Miles travelled per day (average)
- Weekly payment (average)
- Payment variation (average)

  This variable was inspired by the work of Conroy et al., who defined it as "the differences employees experience in their pay from year-to-year, quarter-to-quarter, week-to-week, and day-to-day" [CRDG18]. They found that an increase in variability contributes to drivers turning over. We define *payment variation* as the standard deviation of a driver's average weekly payment, with the hypothesis that a higher standard deviation will positively influence a driver to leave. To the best of our knowledge, we are the only work besides Conroy et al. to include this as a feature variable.

- Trips per day (average)
- On-Duty hours per day (average)
- Age
- Percent of trips with a second driver
- Border crosses per day (average)

- Terminal

  The city base for the driver (e.g. Winnipeg, Calgary, Edmonton).

- Fleet Classification 1

  The type of route the driver took (e.g. City, Highway, Longhaul).

- Fleet Classification 2

  The region in which the driver was based (e.g. West, Central, East)

- Fleet Activity

  The payment type of the driver (e.g. Hourly, Mileage)

- Division

  What division the driver worked in at Company XYZ.

  Division has been anonymized at the request of Company XYZ.

There were other variables considered in early stages of developing the model (such as Marital Status), but all of them were eventually dropped either due to a lack of reliable data or because they did not produce any improvement in the model.

One important decision made at this stage was to separate the data according to driver type. Specifically, we isolated Company drivers and Owner-Operators. A breakdown of these numbers can be seen in Table 4.2. Drivers who were neither

Table 4.2: Driver Breakdown

Type	Left	Retained	Total
Company	405	468	873
Owner-Operator	109	194	303

Company nor Owner-Operator were ignored for this analysis, as we had insufficient data for them.

The main reason for the split was that Company drivers and Owner-Operators are paid differently, and thus required separate aggregation calculations for the Weekly Payment variable. It was also at Company XYZ's request, as they wished to investigate if the model would discover a different set of turnover risk variables for the different types of driver.

4.4 Creating the Model

We chose a Random Forest as our model for this project. As discussed above, Random Forests are straightforward and easy to implement, and they are well-suited to this sort of classification problem. We used the RandomForestClassifier built into Python's scikit library [PVG+11]. For the training and testing step, we used an 80/20 split.

In our main model, we separate the drivers by year of hire, training the model on drivers hired before 2018, and validating it on drivers hired in 2018. We used a 50% binary classification; if the model determined a driver was at least 50% likely to leave, it predicted them to leave, otherwise it predicted them to stay.

In order to optimize our model, we took two steps: hyperparameter optimization and data view optimization. More details on each can be found in the following sections.

4.4.1 Optimizing Hyperparameters

The performance and speed of a Random Forest model can be heavily dependent on the hyperparameters chosen when the model is run. We optimized six different hyperparameters in a two-step process. First, we randomly selected some data and tried a set of possible parameters, which allowed us to narrow down the range of possible values. Second, we ran a grid search on the values from the first step to find the optimal choice for each hyperparameter. All in all, we tested more than 4000 different combinations before we found our ideal set.

The value shown in blue for each hyperparameter is the best value we found in our optimization and the one that was used in our analysis.

4.4.1.1 Hyperparameters

1. Number of Estimators

   This is the number of trees the algorithm will generate. In general, more trees equals better performance but also longer runtime.

   Best value: 1000

2. Bootstrap

   Whether or not samples are drawn with replacement from the original data when constructing the trees.

   Best value: False (no replacement)

3. Min Sample Leaf

   The minimum number of leaves required to split an internal node.

   Best value: 1

4. Min Sample Split

   The minimum number of samples required to split an internal node.

   Best value: 5

5. Max Features

   The maximum number of features the algorithm will place in an individual tree.

   Best value: 11

6. Max Depth

   The maximum depth of the tree.

   Best value: 5

4.4.2 Optimizing Data Views

Once the base model was in place, the other piece we could optimize was the way we fed the model the data we had engineered. As with the hyperparameter optimization, this involved repeatedly testing the model with different options to see which produced the best results. The variables we tested were as follows:

1. Driver Type

   Either Company or Owner Operator drivers

2. Tenure Usage

   Whether or not we used "Years at Company" as a feature variable. Our initial analysis showed this to be an overwhelmingly strong contributor to the model's results (in some cases it accounted for 60% of the weight), so we tested to see how the model would run without it.

3. Tenure Deadline

   A number of days after which a driver would be considered "retained". For example, if this variable was 90, then all drivers who stayed with the company at least three months would be considered "retained", regardless of whether or not they quit after that date. This was included to try and account for the fact that the plurality of drivers leave within the first year, so this variable allowed us to focus on just the drivers that left after a very short tenure.

   We note that models which have this parameter set to an integer value must have False for Tenure Usage, as this value artificially edits tenure and thus may skew results when years-at-company is used in the model.

4. Window Endpoint

   A variable that (with Window Size) allowed us to time-synchronize the data.

5. Window Size

   A variable that (with Window Endpoint) allowed us to time-synchronize the data. By setting Window Endpoint to "Hire Date" and Window Size to 90, for example, we would tell the model to only aggregate the leg data for the first 90 days of a person's tenure.

6. Skip Days

   A number of days at the beginning of a person's tenure to skip when aggregating leg data. This was included to allow us to ignore the period of time when a driver would be in training, and thus not truly driving for Company XYZ.

4.5 Our Evaluation Metric

The method we used to rank our models was a combination of F1-Score and receiver operating characteristic (ROC)-area under the curve (AUC). Specifically, we calculated our metric, E, as follows:

$$E = \sqrt{F1 \times AUC} \tag{4.1}$$

This is a valid calculation because both F1 and ROC-AUC are on the same scale. In both cases:

- The minimum possible value is 0
- The maximum possible value is 1
- A value of 1 indicates a perfect classifier
- A value of 0.5 indicates a classifier effectively operating at random.

Using the harmonic mean of F1 and AUC ensures that the scale is maintained; i.e. all four points listed above are also true of E.

As both F1-Score and AUC have been criticized as imperfect (see Section 2.8), it is our contention that combining the two will yield a more robust metric of evaluation.

We calculated the E value for both the training and the validation for each version of the model, but our ranking system to determine the "best" model looked only at the validation E score.

4.6 Summary

In this chapter, we described the real-life data collected for use in our big data analytics. This could be used for turnover/churn analysis for various industries. In particular, we conducted our turnover/churn analytics on a specific real-life dataset collected from Company XYZ. We outlined two datasets; the initial dataset that we received and the final dataset that was used to train our model. The initial dataset consisted of over 6,500 drivers and provided us with some general insights about turnover at XYZ. The eventual working dataset contained just over 1,000 drivers of two types: Company and Owner-Operator.

We made several observations about our final dataset. First, we had roughly the same number of turnover and retained drivers. Second, the number of turnover drivers grew linearly between 2015 and 2018. Third, the vast majority of drivers who left the company left within one year of being hired.

We describe the final list of feature variables we used: years at company, average miles travelled per day, average weekly payment, average payment deviation, average trips per day, average on-duty hours per day, age, percent of trips with a second driver, average border crosses per day, terminal (e.g. Winnipeg, Calgary), fleet classification 1 (e.g. City, Highway), fleet classification 2 (e.g. East, West), fleet activity (e.g. Hourly, Mileage), and division (internal to XYZ).

We also outline our model, which is the Random Forest built into Python's scikit library. We used it as a 50% binary classifier; if the model gave the driver at least 50% chance to leave, it predicted them to leave. We list the hyperparameters used to improve the model and the different ways we fed the data to the model in order to approach the problem from a series of different angles.

Finally, we explained our evaluation metric, which is a combination of F1-score and receiver operating characteristic (ROC)-area under the curve (AUC).

Chapter 5

Model Validation and Evaluation

5.1 Overview

The analysis shown in this chapter was done specifically on the data from Company XYZ, but we provide a high-level overview of the process, which could be implemented for almost any company.

1. Rank the scores obtained by the Evaluation Metric (see Chapter 4) to determine the "best" model, as well as any other models of interest (based on the data views chosen).

2. Examine the Confusion Matrix and ROC Curve for interesting models to get a better understanding of the score.

   This step is of particular importance as all single-value evaluation metrics have flaws. When a model performs particularly well, it is worth examining the data to ensure that there was not some sort of obvious factor (such as a skewed dataset) which could explain it.

3. Explore the feature variables and their weights provided by the interesting models.

4. Determine cases where the model correctly predicts unintuitive results.

   What counts as unintuitive to a human is subjective and varies depending on industry and the type of data provided. However, a generic definition usable for any dataset is that an unintuitive result is one where many variables go against the average.

   For example, if the average age of employees who turnover is higher than the average of those who remain, and the average salary of employees who turnover is lower than the average of those who remain, then an instance where the model correctly predicted a young, high-wage employee to leave is unintuitive.

5. If possible, test the interesting models for the ability to transfer across types.

   That is, if your employees are divided up into Groups A and B, can a model be trained on Group A and perform well when tested on Group B?

6. Produce Partial Dependence Plots for interesting models to see how variables work in conjunction with one another.

5.2 Best Model

The following is the best-case results we found (for a model trained and tested on the same type of driver). This model evaluated Company drivers, looking at the first 90 days of their tenure (not skipping any days for training). It used years-at-company as a parameter.

The data for this model can be seen in Tables 5.1, 5.2, 5.3 and 5.4, as well as Figures 5.1 and 5.2

Table 5.1: Model Parameters (Best)

Parameter	Value
Driver Type	Company
Tenure Usage	Yes
Tenure Deadline	N/A
Window Endpoint	Hire Date
Window Size	90 days
Skip Days	0

Table 5.2: Model Evaluation (Best)

Metric	Dataset	Value
ROC AUC	Training	0.970
F1 Score	Training	0.910
Evaluation	Training	0.940
ROC AUC	Validation	0.882
F1 Score	Validation	0.639
Evaluation	Validation	0.750

Table 5.3: Confusion Matrix (Best Model - Training)

	Driver Left	Driver Retained
Predicted Left	66	0
Predicted Retained	13	58

Table 5.4: Confusion Matrix (Best Model - Validation)

	Driver Left	Driver Retained
Predicted Left	53	60
Predicted Retained	0	143

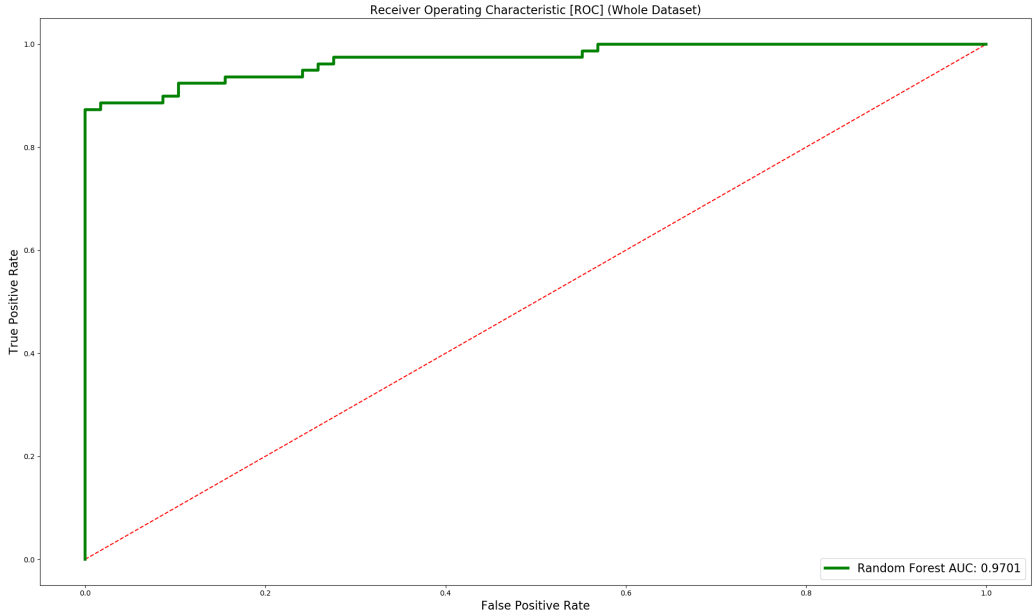

Figure 5.1: ROC for Training Data - Best Result

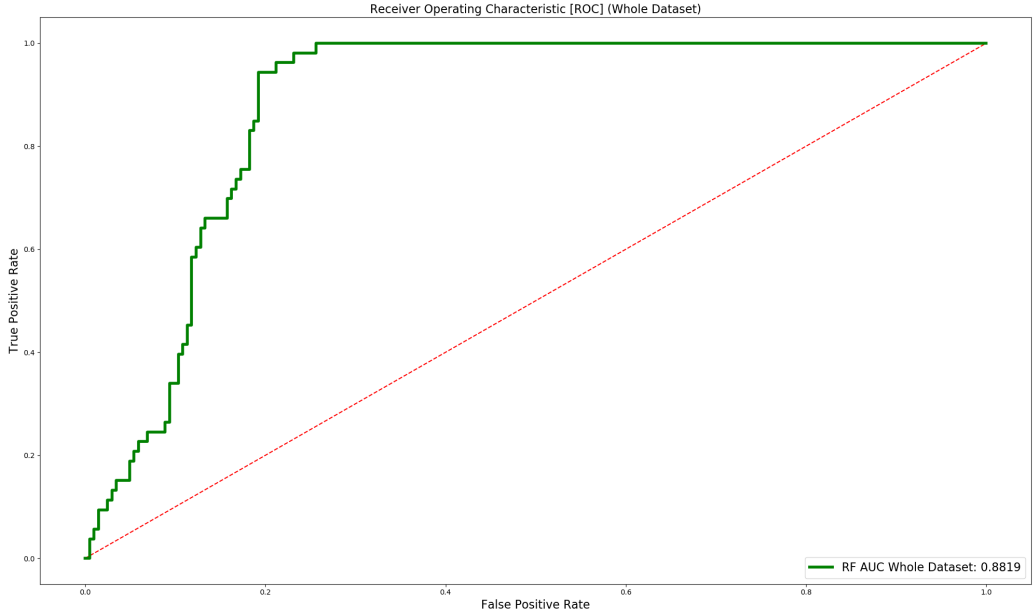

Figure 5.2: ROC for Validation Data - Best Result

5.3 Explaining the Model

Figure 5.3 shows a bar graph displaying how important each of our features was to the model. As mentioned earlier, Years At Company dominates the model, accounting for nearly 80% of the importance. Following it are Average On-Duty Hours Per Day, Average Trips Per Day, and Average Miles Per Day.

Note that the figures shown here are specific to one version of the model; when the model is run with different parameters, the exact weight of each variable will change. However, what is here is a good representation of the general shape of the model.

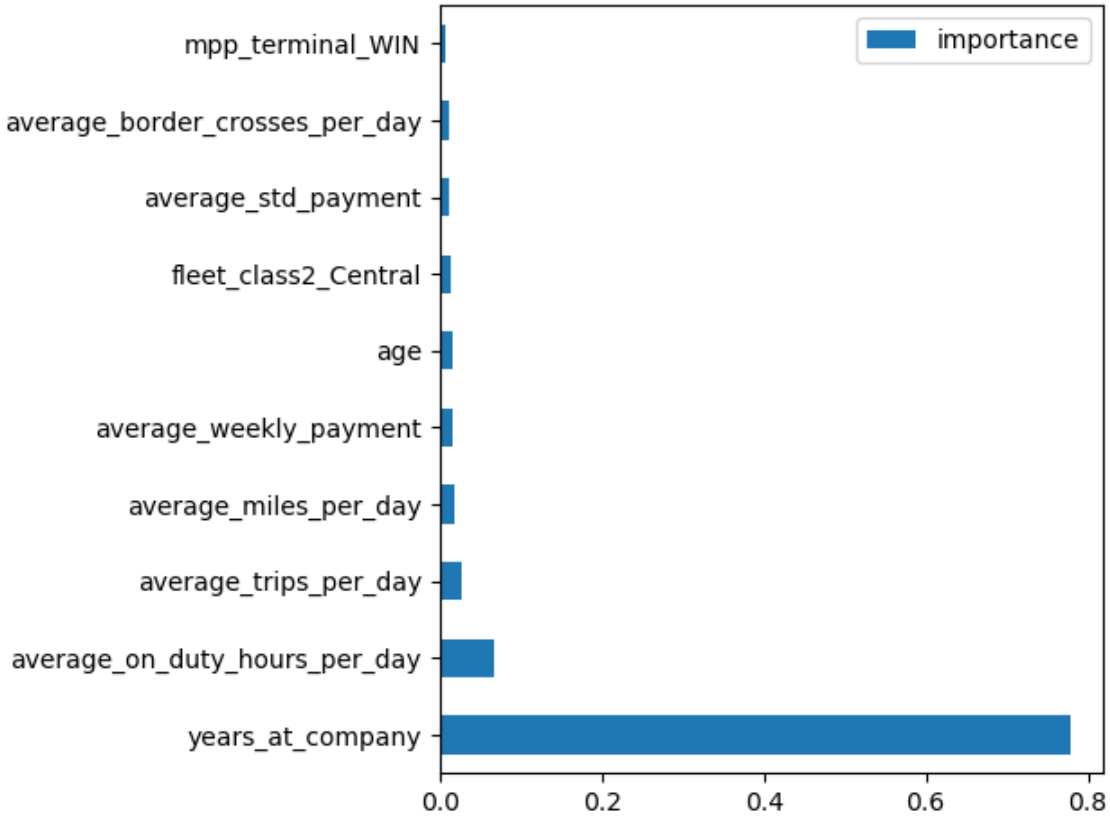

Figure 5.3: Importance of Feature Variables

Figure 5.4 uses SHAP (SHapley Additive exPlanations) [LL17] to explain the model's prediction for an individual driver. The number in bold is the model's prediction (a percentage represented as a decimal) that the driver will leave. In this case, the model predicted this driver would stay at the company (28% likelihood to leave). The blue variables represent things that influenced the model to lower this number (including this driver's 1.29 years at the company), and the size of each section is the amount that each variable influenced the decision. Similarly, the red variables are features which pushed the model toward predicting the driver to leave (including their age of 46).

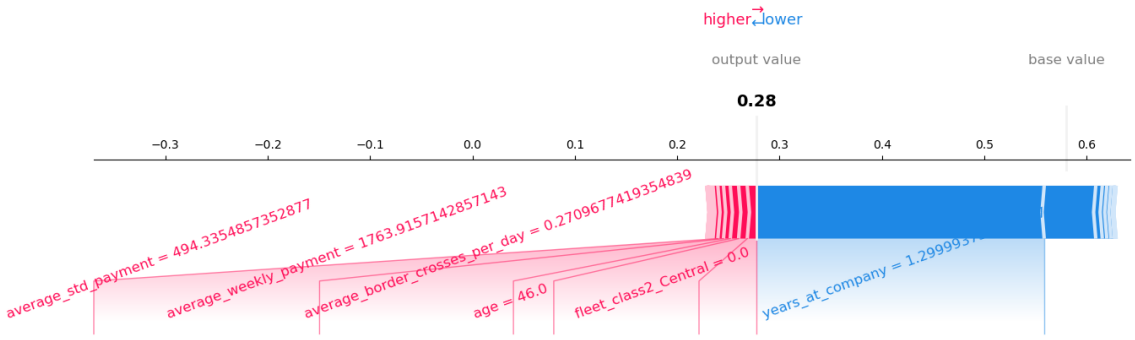

Figure 5.4: Feature Variable Explanation - Single Driver

In Figure 5.5, we can see the model's approach to a second driver. The model predicted this driver to leave (83%), and was pushed to do so by the driver's average on duty hours per day and years at the company, whereas the driver's trips per day pushed that number down.

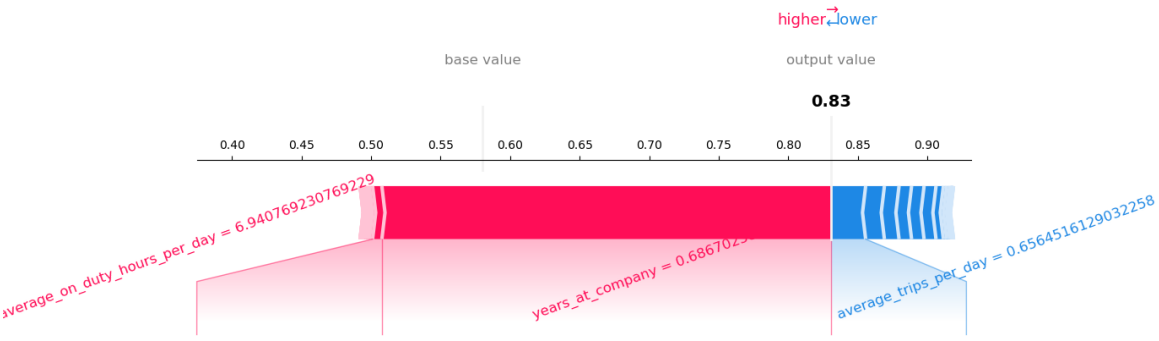

Figure 5.5: Feature Variable Explanation - Single Driver

In Figure 5.6, we can see a broader view of the model. In this image, each point represents an individual driver. The colour of each point represents the relative value of that point's data for that feature (so a pink dot for years-at-company represents an

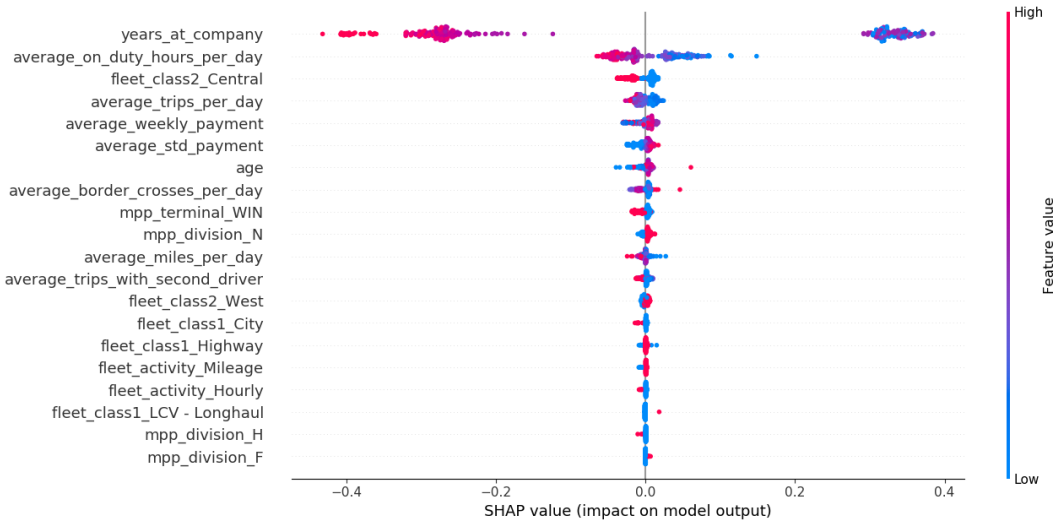

Figure 5.6: Collective Feature Variable Explanation

employee who has been there for a long time) and the left-right position of each point represents how that point's value influenced the model (left for retained, right for turned over). The distance from the y-axis represents the strength of the influence.

This summary chart allows us to make some generalizations about the way that the feature variables influence the model. Years-at-company, for instance, seems to strongly push drivers to leave when it is low (extreme left side is heavily blue) and strongly push drivers to stay when it is high.

More analysis on this version of the model and others can be found in the Discussion section of this chapter.

It is somewhat tempting from a business perspective to look at Figure 5.6, write down a list of each influential factor (as indeed we did) and the direction it influences, and think that can draw all the conclusions. However, knowing the factors in isolation is not enough to replicate the work of the model. In the next section, we discuss a more complicated result based on some non-trivial cases the model predicted correctly.

5.4 Non-Trivial Cases

In Table 5.5, we have the generalized data for several of our factors (in the best model). The second column shows the average value for that factor. The third column is "Retained" if a **high** value for that factor influences the driver to stay with the company (and thus a low value suggests they will leave), and "Turnover" if the reverse is true.

Table 5.5: Feature Averages

Factor	Average	High Influence
Terminal CAL	0.30	Turnover
Terminal EDM	0.17	Turnover
Terminal KEL	0.00	Turnover
Terminal LAN	0.03	Turnover
Terminal MIS	0.13	Turnover
Terminal MTL	0.02	Turnover
Terminal REG	0.05	Retained
Terminal WIN	0.30	Retained
City (FC1)	0.04	Retained
Highway (FC1)	0.95	Turnover
Longhaul (FC1)	0.00	Retained
Central (FC2)	0.35	Retained
East (FC2)	0.14	Turnover
West (FC2)	0.50	Turnover
Hourly (FA)	0.04	Retained
Mileage (FA)	0.95	Turnover
Division B	0.02	Retained
Division F	0.09	Turnover
Division G	0.17	Turnover
Division W	0.01	Turnover
Division N	0.68	Retained
Division H	0.02	Turnover
Age	41.19	Turnover
Years at company	0.81	Retained
Average trips per day	0.58	Retained
Average miles per day	161.23	Retained
Average on duty hours per day	7.80	Retained
Average border crosses per day	0.06	Retained
Average trips with second driver	0.26	Retained
Average weekly payment	1289.14	Retained
Average payment deviation	339.52	Turnover

In Table 5.6, we can see five drivers, all of whom were correctly predicted by our model. The values that have been **bolded** are all values which would "disagree" with our averages table. For example, Driver A is 38 years old. This is under the

Table 5.6: Non-Trivial Cases

Feature	Driver A	Driver B	Driver C	Driver D	Driver E
Left	Y	N	N	Y	N
Age	**38**	38	**67**	**30**	**52**
Years at company	0.22	**0.78**	1.11	0.15	1.15
Avg trips per day	0.10	**0.53**	1.26	0.23	**0.25**
Avg miles per day	51.75	**92.28**	260.71	94.60	**118.50**
Avg on duty hrs/day	4.97	**7.16**	11.31	3.45	10.39
Avg border crosses/day	**0.08**	**0.01**	**0.00**	**0.08**	**0.02**
Avg trips w/ 2nd driver	**1.00**	**0.02**	**0.00**	**1.00**	0.35
Avg weekly payment	862.94	**1189.19**	1290.77	861.11	1295.02
Avg payment deviation	**3.71**	259.86	**401.17**	**0.00**	**355.71**
Terminal	**WIN**	**EDM**	**CAL**	**WIN**	**CAL**
Fleet Class 1	HWY	**HWY**	**HWY**	HWY	**HWY**
Fleet Activity	MLG	**MLG**	**MLG**	MLG	**MLG**
Fleet Class 2	**Central**	**West**	**West**	**Central**	**West**
Division	**N**	N	**G**	**N**	N

average (41), and our Influence column says that a young age value causes drivers be retained. However, Driver A was turned over.

Even our most influential factor, years-at-company, is not immune to being non-trivial. Driver B is below the mean, which our model says encourages turnover, and yet Driver B was retained.

Note that for the Terminal, Fleet, and Division fields, a driver is considered to have a 1 if their Terminal/Fleet/Division matches and a 0 if they do not. Also note that to conserve space, a Fleet Class 1 of "Highway" has been abbreviated to "H" and similarly, a Fleet Activity of "Mileage" has been abbreviated to "M".

The sheer number of times where the unintuitive is the case, just for these five drivers, is a strong indication that the feature variables cannot be looked at in isolation; it is their combination in the model that results in its predictive power.

5.5 Transfer Learning

In our initial runs of the model, we pulled both training and testing data from the same type of driver. In our best-case scenario above, we trained it on Company drivers hired between 2015 and 2017 (inclusive) and tested it on Company drivers hired in 2018. However, in order to demonstrate the model's predictive power, we also ran it across driver types. Specifically, we trained it on Company drivers (2015-2018) and tested it on Owner-Operator drivers (2015-2018).

Several transfer models actually outperformed the model shown above; the data shown here is for the best of those. Our supposition is that the increased amount of data available (for both training and validation) allowed the model to improve in the transfer case.

Note that when this thesis refers to the "Best Model", it is exclusively referring to the model from Section 5.2, even though the E score for this model was higher. "Best" in this case can be thought of as a shorthand for "Best of the original models", all of which were trained and tested on the same type of driver.

The data for this model can be seen in Tables 5.7, 5.8, 5.9 and 5.10, as well as Figures 5.7, 5.8 and 5.9.

The SHAP Summary graph for this model (Figure 5.9) is particularly interesting as it is not very symmetric. In the years at company field, for instance, we can see that the blue dots extend much farther to the right than the red dots do to the left, and there are also some blue dots to the left of the y-axis. This implies that a driver's tenure either pushed the model very strongly toward leaving or somewhat strongly toward staying. Similar asymmetric distributions happen for almost every variable,

Table 5.7: Model Parameters (Transfer)

Parameter	Value
Driver Type (Training)	Company
Driver Type (Testing)	Owner-Operators
Tenure Usage	Yes
Tenure Deadline	None
Window Endpoint	Termination Date
Window Size	90
Skip Days	0

Table 5.8: Model Evaluation (Transfer)

Metric	Dataset	Value
ROC AUC	Training	0.951
F1 Score	Training	0.847
Evaluation	Training	0.898
ROC AUC	Validation	0.867
F1 Score	Validation	0.769
Evaluation	Validation	0.816

Table 5.9: Confusion Matrix (Transfer - Training)

	Driver Left	Driver Retained
Predicted Left	61	2
Predicted Retained	20	92

Table 5.10: Confusion Matrix (Transfer - Validation)

	Driver Left	Driver Retained
Predicted Left	88	32
Predicted Retained	21	162

with on duty hours per day and average payment variance standing out as extending in virtually only one direction.

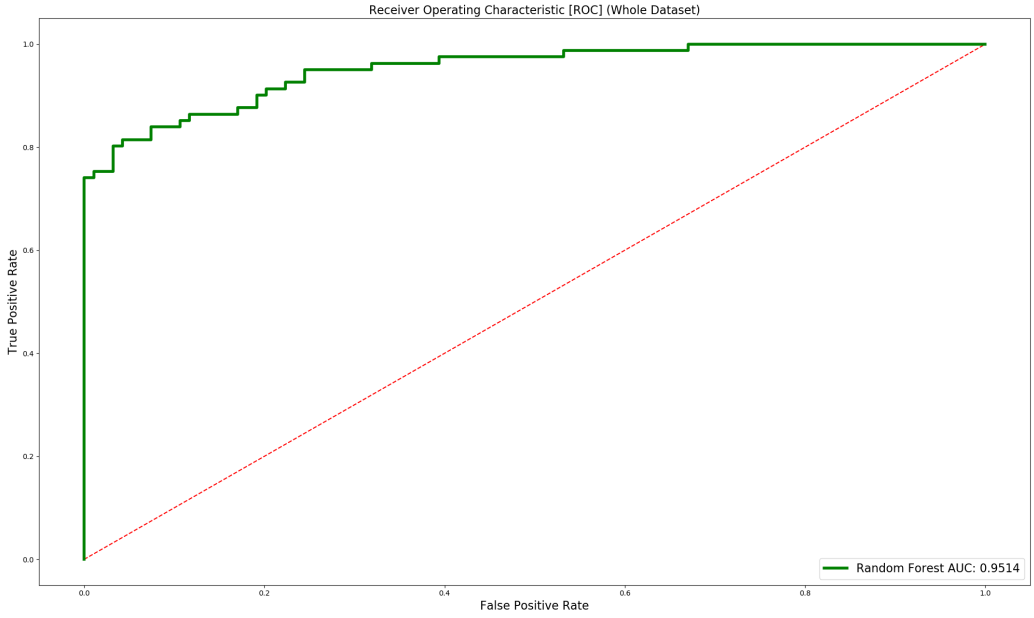

Figure 5.7: ROC for Training Data - Transfer

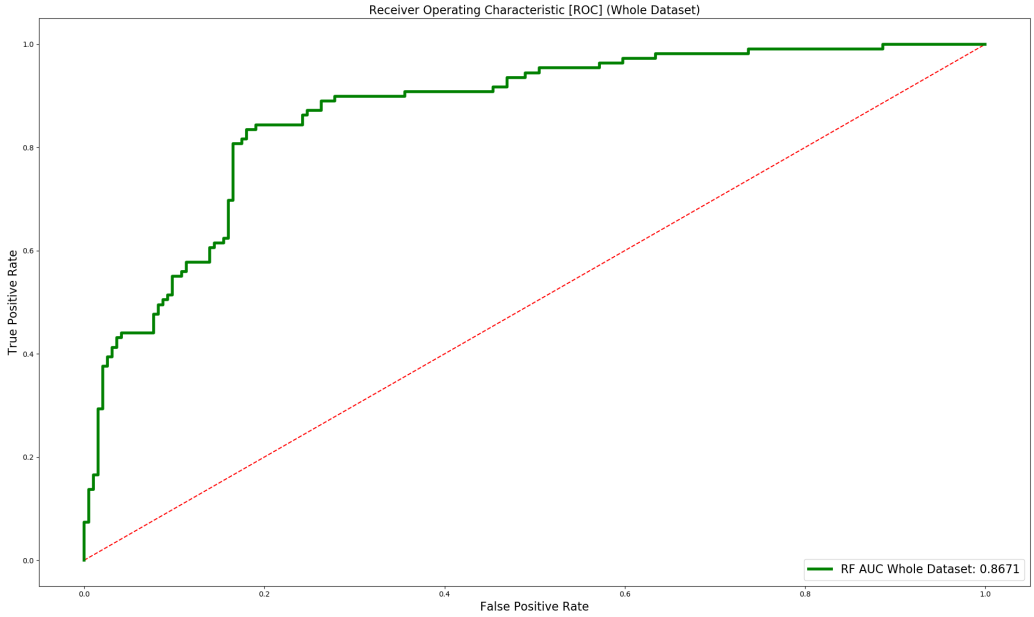

Figure 5.8: ROC for Validation Data - Transfer

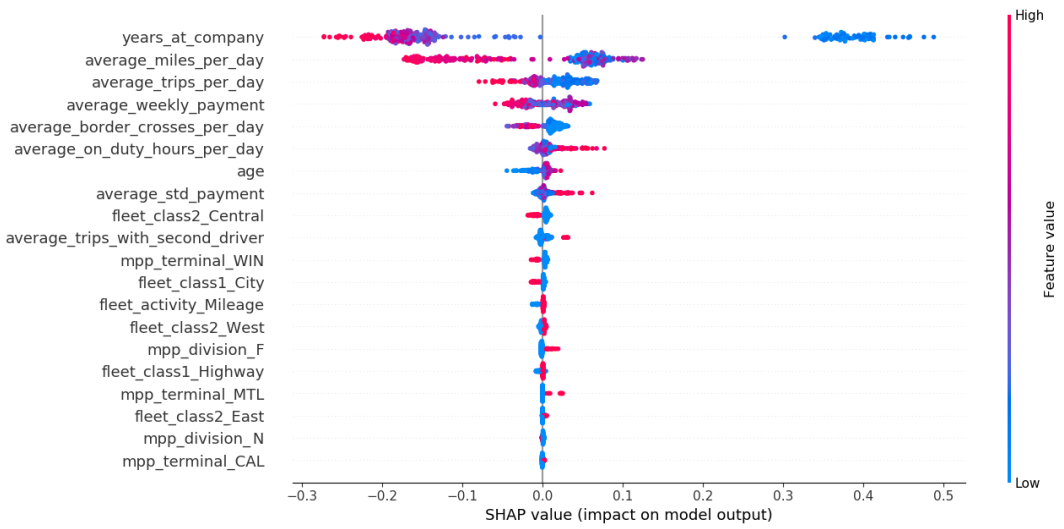

Figure 5.9: SHAP Summary - Transfer

5.6 Without Years at Company

The fact that the feature years-at-company dominates the model so strongly led us to be curious as to how well the model could perform without it. As mentioned above, one of the variables we tweaked when optimizing the model was whether or not to include years-at-company in the analysis.

Models that ignored years-at-company did about as well on average as models that included it (we examine this in further detail below). The results of the best model ignoring years-at-company are shown below, which happens to be the second-best model overall, coming in just behind the transfer model shown in the previous section. It is likely not a coincidence that this model has identical parameters to the best transfer model apart from not using years-at-company.

The data for this model can be seen in Tables 5.11, 5.12, 5.13 and 5.14, as well as Figures 5.10, 5.11 and 5.12

Table 5.11: Model Parameters (No Years At Company)

Parameter	Value
Driver Type (Training)	Company
Driver Type (Testing)	Owner-Operators
Tenure Usage	No
Tenure Deadline	None
Window Endpoint	Termination Date
Window Size	90
Skip Days	0

Table 5.12: Model Evaluation (No Years At Company)

Metric	Dataset	Value
ROC AUC	Training	0.791
F1 Score	Training	0.720
Evaluation	Training	0.755
ROC AUC	Validation	0.858
F1 Score	Validation	0.697
Evaluation	Validation	0.773

Table 5.13: Confusion Matrix (No Years At Company - Training)

Prediction	Left	Retained
Retained	63	31
Left	18	63

Table 5.14: Confusion Matrix (No Years At Company - Validation)

Prediction	Left	Retained
Retained	92	63
Left	17	131

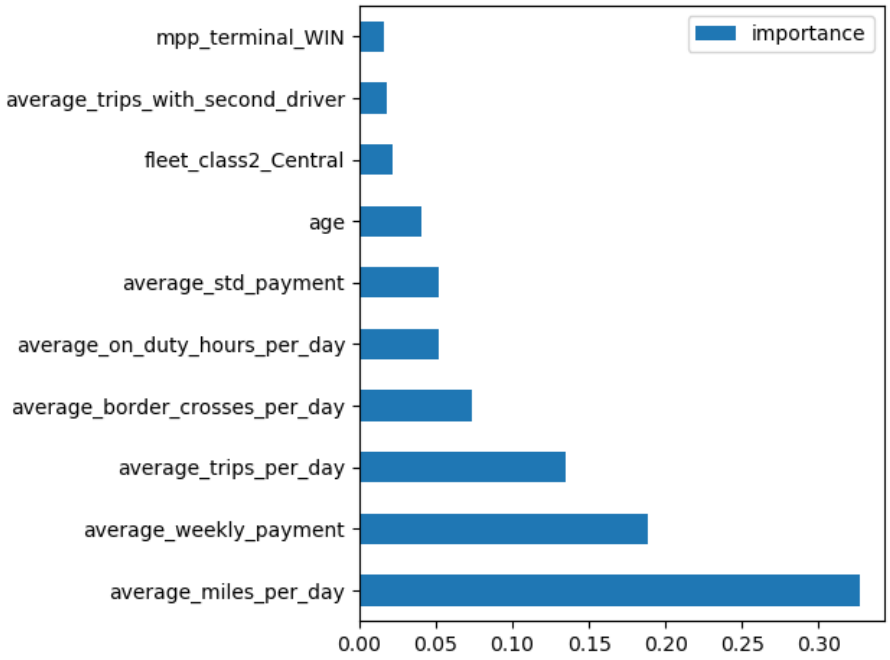

Figure 5.10: Feature Importance, No Years at Company

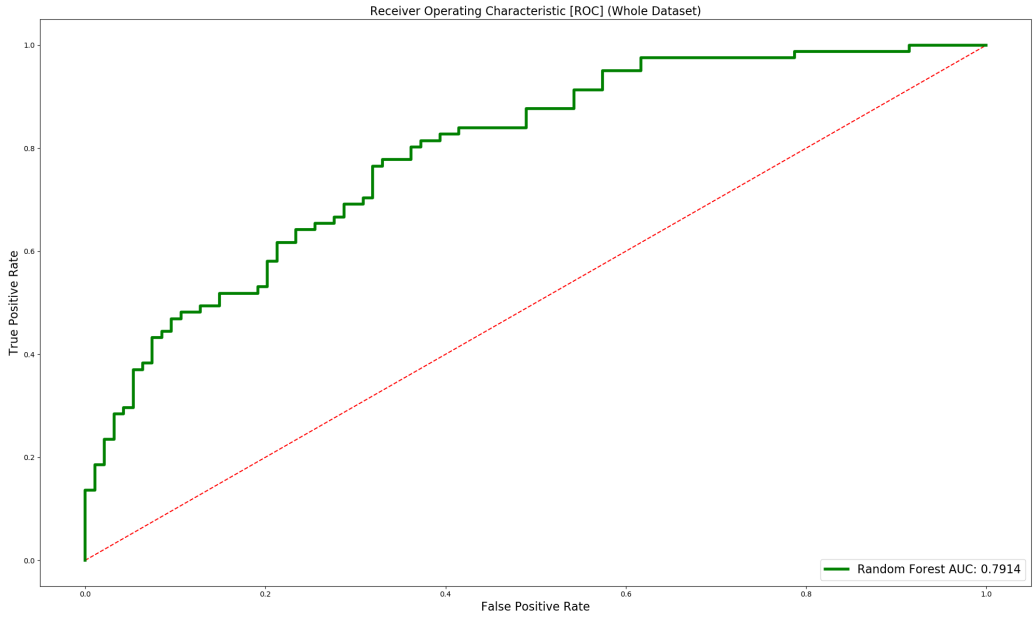

Figure 5.11: ROC for Training Data - No Years At Company

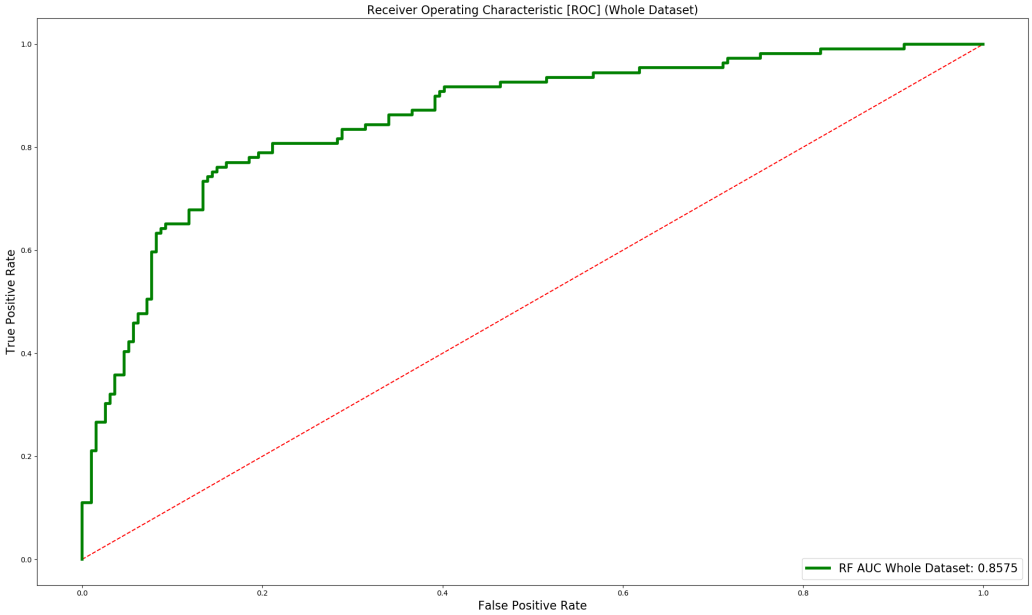

Figure 5.12: ROC for Validation Data - No Years At Company

5.7 Discussion

Across the more than 70 models we tested, we noticed some interesting patterns.

5.7.1 Important Variables

For each iteration of our model, we logged the ten most important feature variables, which accounted for an average of 95% of the results. In Table 5.15, the Frequency column shows the percentage of results in which each feature variable was in the top ten. A visual representation of the same data is shown in Figure 5.13. In the Average column, we show the average explaining power of each feature variable *using only the cases where it was in the top ten.* In the Zero Avg column, we show the average explaining power of each feature variable *giving it a score of zero when*

Table 5.15: Average Variable Importance

Feature	Frequency	Average	Zero Avg
Years at Company	34%	61.0%	20.9%
Avg Trips Per Day	100%	15.3%	15.3%
Avg Miles Per Day	100%	14.2%	14.2%
Avg On-Duty Hours Per Day	100%	11.2%	11.2%
Avg Weekly Payment	100%	9.4%	9.4%
Avg Payment Deviation	100%	7.3%	7.3%
G Division	21%	3.8%	0.8%
Age	99%	3.6%	3.6%
EDM (Terminal)	3%	3.5%	0.1%
Avg Border Crosses Per Day	86%	3.4%	2.9%
Avg Trips with Second Driver	63%	3.2%	2.0%
F Division	38%	3.1%	1.2%
N Division	21%	2.9%	0.6%
Mileage (FA)	16%	2.5%	0.4%
Highway (FC1)	26%	2.4%	0.6%
Hourly (FA)	7%	2.4%	0.2%
Central (FC2)	52%	2.3%	1.2%
WIN (Terminal)	19%	2.2%	0.4%
LAN (Terminal)	3%	2.1%	0.1%
West (FC2)	15%	1.9%	0.3%
City (FC1)	5%	1.7%	0.1%
MIS (Terminal)	3%	1.5%	0.0%
CAL (Terminal)	7%	1.1%	0.1%

it was not in the top ten. FC1, FC2, and FA stand for Fleet Classification 1, Fleet Classification 2, and Fleet Activity, respectively.

Note that Years At Company is skewed in both the Frequency and Zero Avg columns as one of our variable parameters was toggling whether or not it was even being considered.

From both the table and the chart, six factors stand out clearly as the most influential.

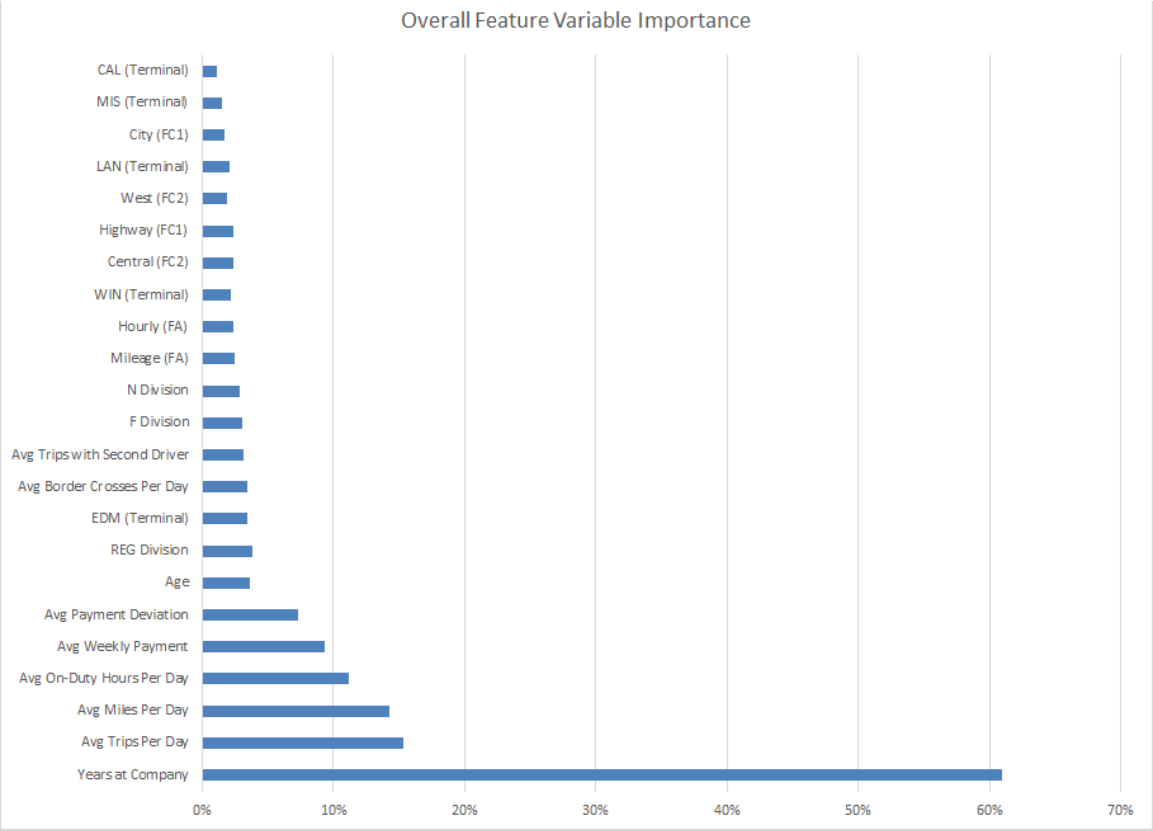

Figure 5.13: Importance of Feature Variables, Overall

- Years worked at the company
- Miles travelled per day
- Trips taken per day
- Average weekly payment
- Average payment deviation
- Average on-duty hours per day

Each of these variables shows up in the top ten list on every model (years-at-

company shows up in 100% of the models in which it is used as a feature), and there is a marked split between Average Payment Deviation and the features below it in the table; everything from REG downward has an average of less than 4%.

One item of note about these variables is that they are all independent of the company. Certain feature variables on our full list, most notably Division but also the two Fleet Class Variables and Fleet Activity, were either internal to Company XYZ or categorized by XYZ. In either case, while some other company may track the same data, they may not describe it in the same way, which would limit the transferability of our model. The six variables shown here, fortunately, do not have that limitation, and can be applied to any company's data.

5.7.2 Partial Dependency Plots

Unfortunately, the nature of Random Forests is such that we cannot simply choose a variable we know to be important (say, Average Trips per Day) and then answer the question of whether having a high number of Average Trips per Day makes a driver leave or stay. Even if we could, it would not yield reliable results (see the section on Non-Trivial Cases for an example of this). Still, plots which depict the effects of individual variables are useful visualizations.

We produced partial dependence plots (PDPs) for the five most influential feature variables for each model, as well as the two-way PDP for every possible combination of those top five variables. In this section, we show some PDPs for our Best Model (Section 5.2).

In Figure 5.14, we can see the PDP for Years at Company in our best model.

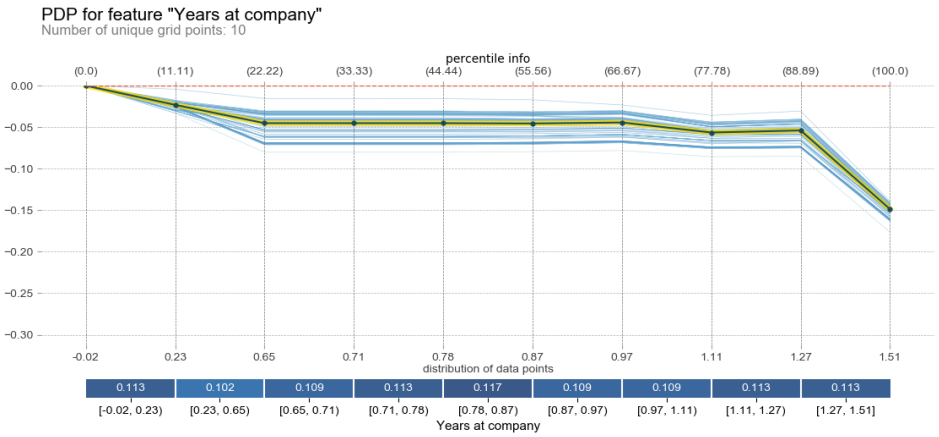

Figure 5.14: 1D Partial Dependence Plot for Years At Company - Best Model

The Y-axis of the PDP represents the *relative* change in the model's prediction as the feature value increases. In this case, as Years at Company increases, the model becomes more likely to predict the negative case (i.e. that the driver will be retained). However, there is little change in the model for YAC values between 0.65 and 0.97.

In Figure 5.15, we see a different PDP, showing Trips per day (TPD), using on-duty hours per day as a reference colouring. The plot shows that when the number of trips per day is low, the on-duty hours per day are a bigger factor (with low numbers causing the model to more strongly predict a driver to leave).

We can see another view of Trips per day in Figure 5.16. In this image, we can see that TPD has an interesting trend. For the lowest 10% of data, the model's prediction score increases, but then it drops down to just under the baseline for the majority of TPD values, only to spike up again for the last 20% of data. This suggests that there is a "sweet spot" of Trips per day that drivers like; having too many or too few

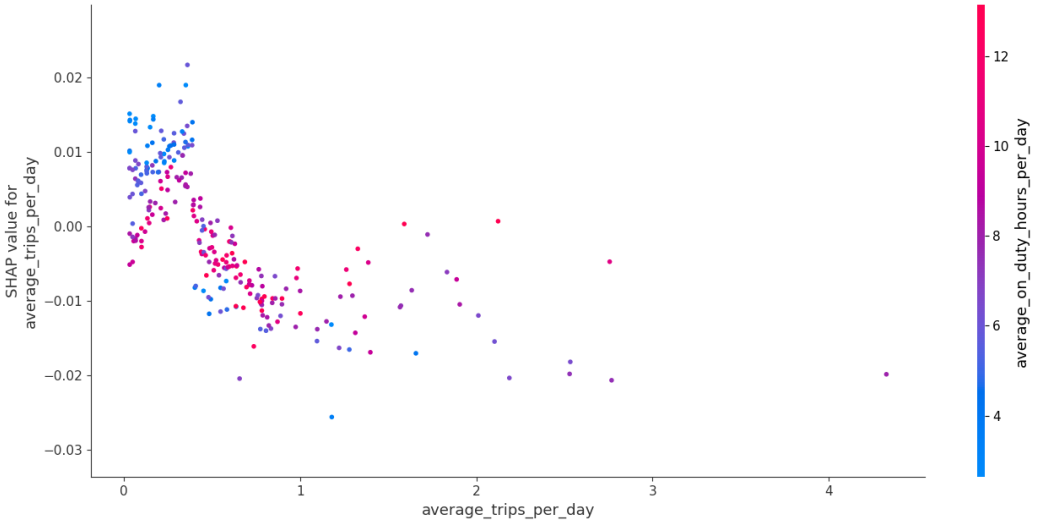

Figure 5.15: Partial Dependence Plot for Trips Per Day - Best Model

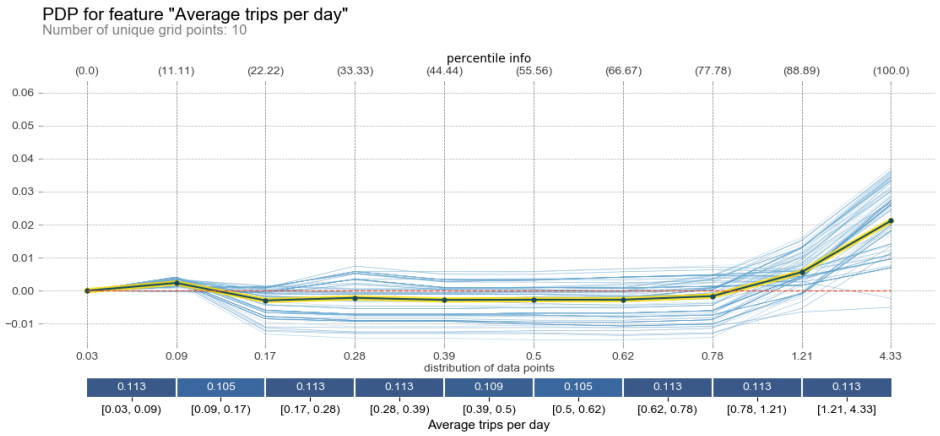

Figure 5.16: 1D Partial Dependence Plot for Trips Per Day - Best Model

(especially too many) encourages them to leave.

In Figure 5.17, we see an interesting PDP, showing on-duty hours per day, using WIN Terminal as a reference colouring. The plot shows a general downward trend

in the model's prediction as the on-duty hours increase, but with a pivot across WIN. When on-duty hours are under 7, being based in the WIN terminal increases a driver's probability to leave. However, when on-duty hours are greater than seven, the influence of WIN flips, and it decreases the leaving probability.

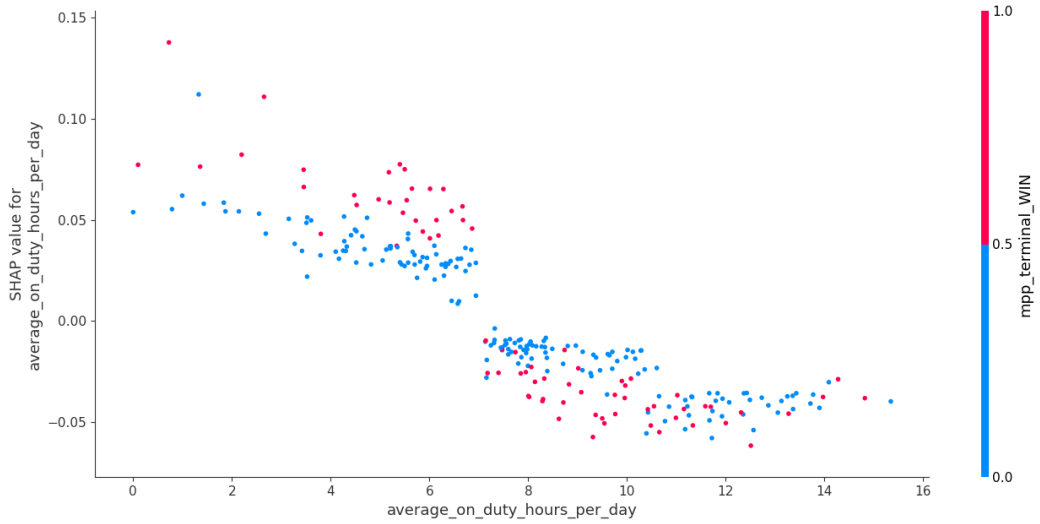

Figure 5.17: Partial Dependence Plot for On Duty Hours - Best Model

Figure 5.18 shows a 2D contour PDP, plotting years at company against average trips per day, with the colouring representing the model's prediction score. In large part, we see that years at company's influence on the model is independent of trips per day, but there are notable sections where this is not true. When years at company is close to zero, for example, a very high or very low number of trips per day increases the model's prediction score (this matches up with our conclusion earlier that drivers do not like having too many or too few trips per day). Similarly, for drivers who have been at the company between 1 and 1.3 years, having trips per day be above one raises the model's prediction value.

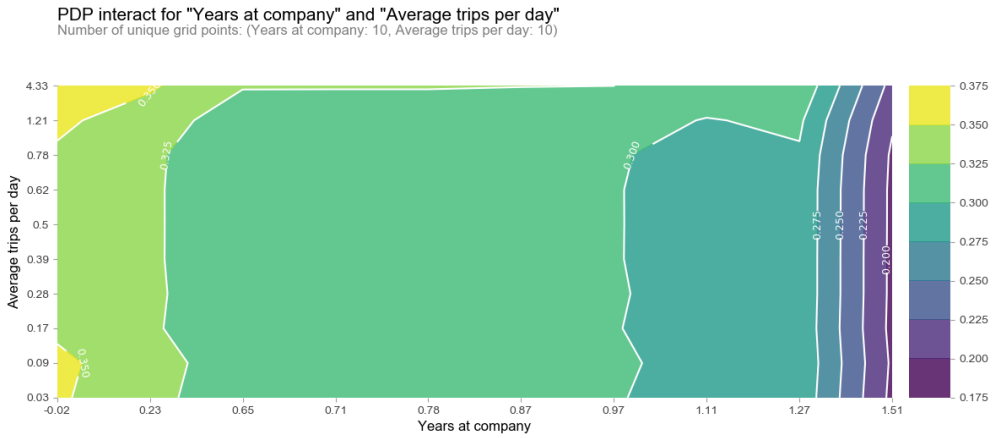

Figure 5.18: 2D PDP for Years at Company and Trips Per Day - Best Model

To see more strongly connected variables, we look at Figure 5.19. Here we see an interesting exchange. Once on-duty hours (ODH) per day gets above 4.26, the model only cares about average trips per day, having its lowest prediction between 0.17 and

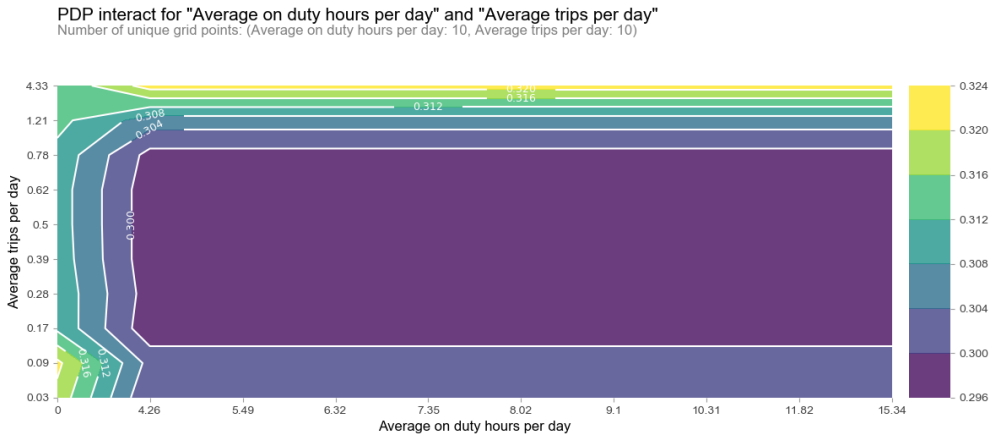

Figure 5.19: 2D PDP for On-Duty Hours and Trips Per Day - Best Model

0.78 TPD and then increasing at the top end. Below 4.26 on-duty hours per day, though, the model is much more concerned with ODH; TPD is almost a non-factor except at the extreme top and bottom ends.

5.7.3 Other PDPs

In this section, we feature PDPs for the transfer model outlined in section 5.5 and the No-YAC model described in section 5.6.

5.7.3.1 Transfer PDPs

In Figure 5.20, we can see three distinct layers for years at company; from 0 to 9 months, from 9 months to 2 years, and 2+ years. Consistent with our original hypothesis and previous results, predictions to leave decrease as years at company goes up.

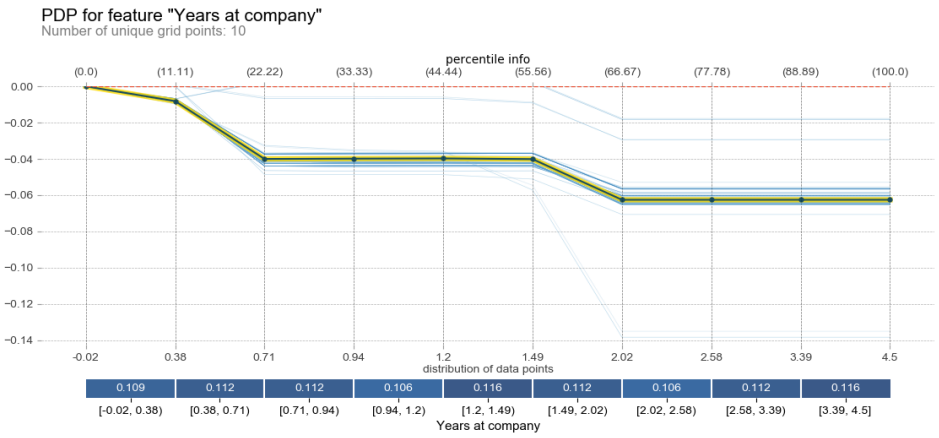

Figure 5.20: 1D Partial Dependence Plot for Years At Company - Transfer Model

In Figure 5.21, we see a different shape than we saw in our other model for trips per day. However, we note the differing scales on the X-axis. In the model shown in section 5.2, 90% of our data points were below 1.21 TPD, whereas this dataset has a much broader scope.

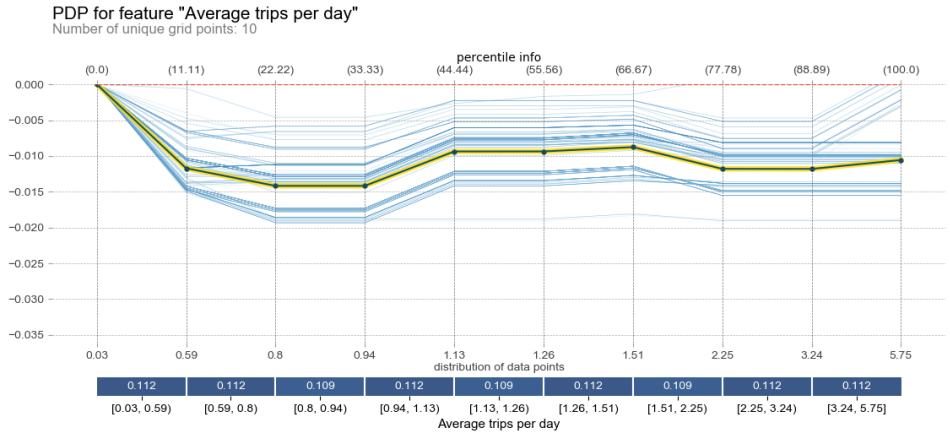

Figure 5.21: 1D Partial Dependence Plot for Trips Per Day - Transfer Model

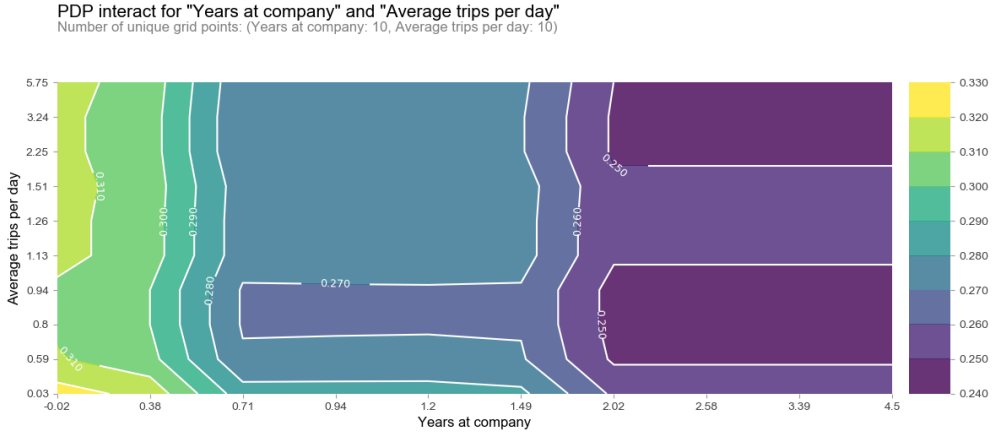

Figure 5.22: 2D PDP for Years at Company and Trips Per Day - Transfer Model

5.7.3.2 No YAC PDPs

With no years-at-company, we include the PDPs for the top three feature variables, which are much more evenly distributed than in models using YAC.

We include Figure 5.23 as a curiosity; the model lists average weekly payment as the second most important variable, with just under 20% importance, and yet its PDP is a straight line on the baseline. At present, we have no explanation for this.

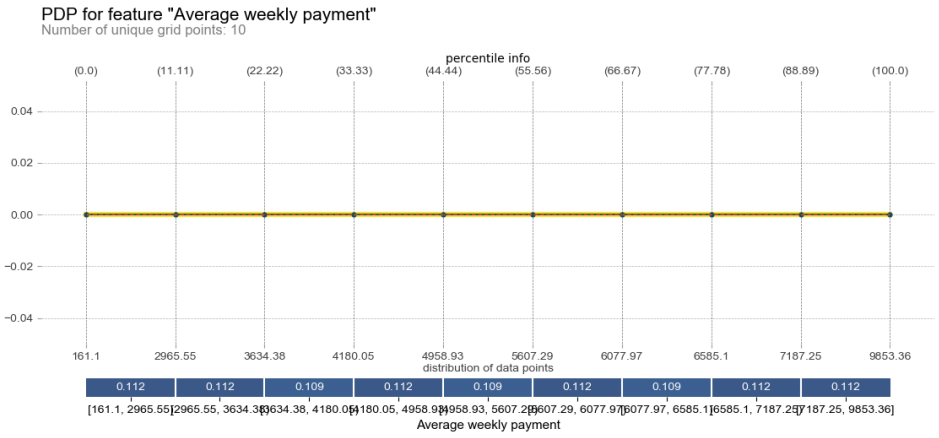

Figure 5.23: 1D Partial Dependence Plot for Weekly Payment - No YAC Model

We can see the impact of Average Weekly Payment more clearly in Figure 5.24, where there appears to be a pivot point at about 6000. We also note a curious group of drivers in the top-left of that graph, who appear to have a high number of miles per day (from the colouring) and yet have the lowest average weekly payment. Unsurprisingly, such drivers appear to be more prone to leaving.

Of the models on display in this chapter, only this one has both Average Weekly Payment and Average Payment Deviation in its top five feature variables, and Fig-

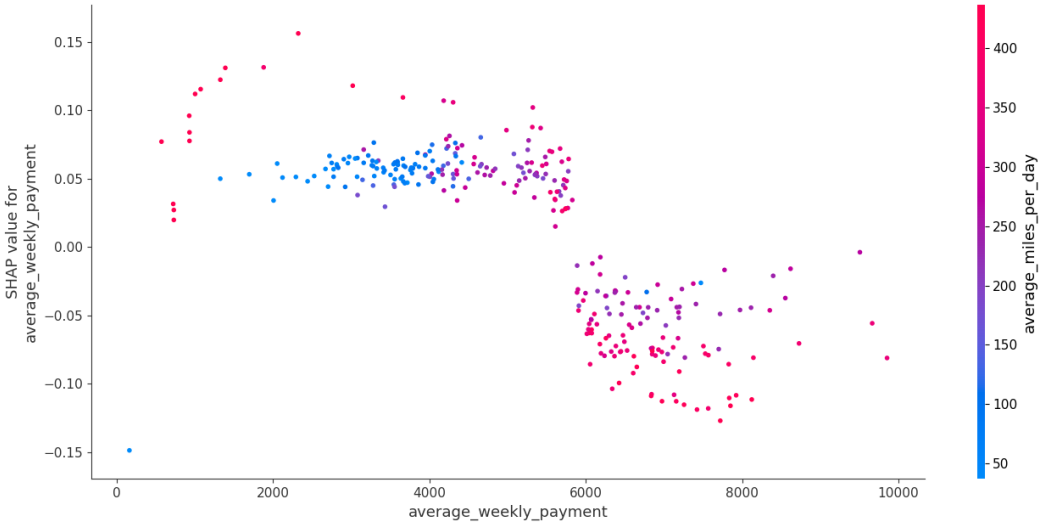

Figure 5.24: Partial Dependence Plot for Average Weekly Payment - No YAC Model

ure 5.25 shows the PDP of their interactions. Perhaps surprisingly, there is no association shown; Average Payment Deviation appears independent of Average Weekly Payment, and beyond a deviation of 807, it does not appear to influence the model further.

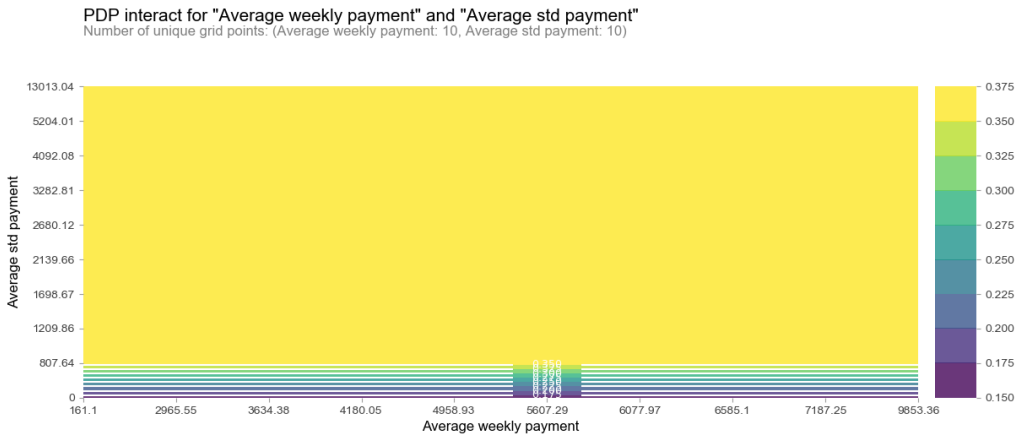

Figure 5.25: 2D PDP for Years at Company and Trips Per Day - No YAC Model

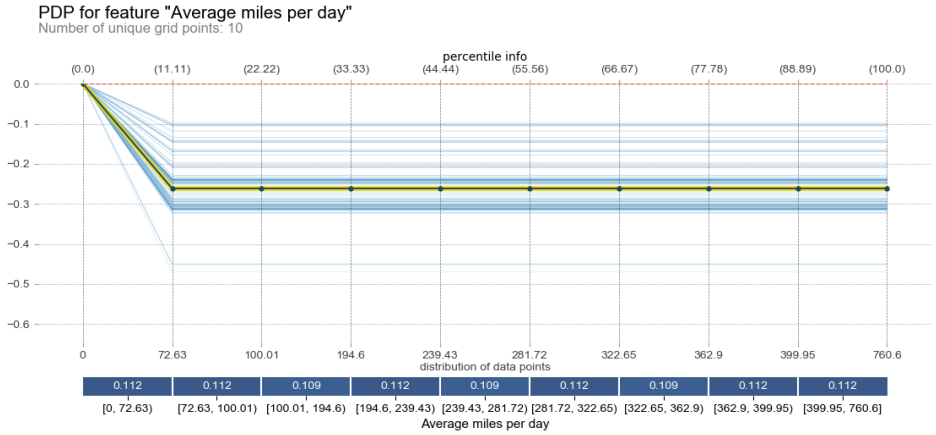

Figure 5.26: 1D Partial Dependence Plot for Miles Per Day - No YAC Model

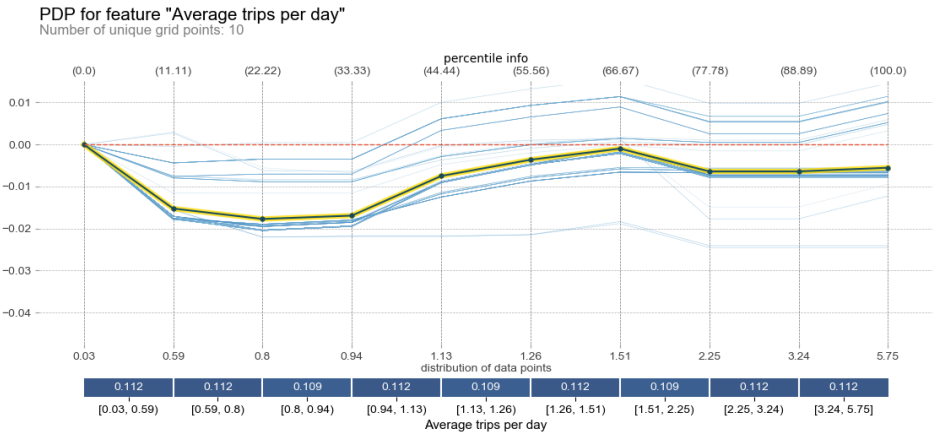

Figure 5.27: 1D Partial Dependence Plot for Trips Per Day - No YAC Model

5.7.4 Connection between Model Changes and Important Variables

Our six key features identified earlier show up in every model, but the other four entries on the top ten list for each model varied quite a bit. In this section we discuss the relationship between the model parameters we chose and the feature variables that were important to that model. Specifically, we look at the six variables which showed up in the top ten list in at least 25% of the models: Age, Average Border Crosses Per Day (ABC), Average Trips with a Second Driver (ATS), Central FC2 and Highway FC1. We ignore R Division even though it showed up in 38% of models as it is an internal label for Company XYZ and may have explanatory factors for these correlations we do not know.

5.7.4.1 Age

Age appeared in the top ten list for nearly every model, but its overall importance averaged below 4%. It did not have any strong correlations with changing the model parameters.

5.7.4.2 Average Border Crosses Per Day

ABC appeared in the top ten list for 86% of models. It was most strongly connected to tenure (appearing in 92% of models where years-at-company was not used as a feature variable, but only 75% of models where it was) and training data (appearing in 97% of models trained on company drivers, but only 75% of those trained on owner-operators).

5.7.4.3 Average Trips with a Second Driver

ATS appeared in 63% of models. Its appearance was correlated with several model parameters.

- Tenure Usage: 69% of models that ignored tenure, 50% of models that used it.
- Max Tenure: 75% of models that used a 360-day cutoff, 57% of models that did not.
- Training Data: 86% of models trained on company drivers, 39% of models trained on owner-operators.
- Window Size: 92% of models that looked at a driver's entire tenure, 48% of models that used a 90-day window.

5.7.4.4 Fleet Class 2 Central

Central appeared in 52% of models (West and East appeared in only 15% and 0% of models, respectively). Its appearance was correlated with several model parameters.

- Days Skipped: 64% of models that did not skip days, 39% of models that did.
- Max Tenure: 61% of models that had no cutoff, 33% of models that had a 360-day cutoff.
- Window Endpoint: 67% of models that started from termination date, 42% of models that started from hire date.

- Training Data: 61% of models trained on company drivers, 42% of models trained on owner-operators.

5.7.4.5 Fleet Class 1 Highway

Highway appeared in 26% of models (City and Longhaul appeared in only 5% and 0% of models, respectively). Its appearance was correlated with several model parameters.

- Tenure Usage: 31% of models that used tenure, 13% of models that ignored it.
- Max Tenure: 42% of models that used a 360-day cutoff, 18% of models that did not.
- Window Endpoint: 58% of models that started from termination date, 4% of models that started from hire date.
- Window Size: 31% of models that used a 90-day window, 13% that looked at a driver's entire tenure.

5.7.5 Years at Company

As is obvious from both Figures 4.2 and 5.3, the years an employee has spent working at Company XYZ is a very significant factor when it comes to the decision to leave. As was mentioned earlier, on average models did roughly equally well when they included or excluded years-at-company, but breaking this down reveals that they actually do much worse without it, except in a specific case.

Table 5.16: Average E Score with and without Years At Company

Trained	CO	OO	CO	OO
Tested	CO	OO	OO	CO
YAC Ignored	0.52	0.39	0.645	0.575
YAC Used	0.70	0.31	0.743	0.742

Table 5.17: Maximum E Score with and without Years At Company

Trained	CO	OO	CO	OO
Tested	CO	OO	OO	CO
YAC Ignored	0.58	0.55	0.773	0.686
YAC Used	0.75	0.34	0.816	0.766

In Tables 5.16 and 5.17, we can see the pattern; having years-at-company is beneficial to most types of model, but detrimental to models trained and tested on Owner-Operators.

One possible explanation for this is the extreme skew in data we have for the validation set in that case. There are 78 drivers who were retained and only 2 who left, so it is likely that the departure from the pattern is due to the nature of the data and not a fundamental difference in the way that Owner-Operators work.

5.8 Summary

In this section we described the results of some of our models, including the best one overall (a transfer model), the best one trained on and tested on the same type of driver, and the best one that deliberately ignored years-at-company as a feature variable.

We provided several visuals to help explain the impact of the feature variables

on each model we chose and talked about the interpretations of those visuals. We also discussed correlations between our model parameters and which feature variables were deemed important by the model.

We identified six variables which were consistently important to the model across all variations (in decreasing order of average importance):

- Years worked at the company
- Miles travelled per day
- Trips taken per day
- Average on-duty hours per day
- Average weekly payment
- Average payment deviation

From the SHAP Summary (Figure 5.6) of our best non-transfer model, as well as partial dependence plots, we conclude that in general a driver who leaves has some combination of the following traits:

- Low-to-average years worked at the company
- Low miles travelled per day
- An extreme (low or high) number of trips taken per day
- Under seven on-duty hours per day
- Low weekly payment

- High payment deviation

Finally, we also discussed how removing years-at-company as a feature impacted the score of the model. We found that most types of models perform worse without out, but models trained and tested on owner-operators perform better. We speculated that this had more to do with the nature of the owner-operator validation data than years-at-company itself.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The goal of this thesis was to discover what combination of features, data, and parameters would yield the most accurate prediction model for driver turnover. Existing work in this field, to the best of our knowledge, as focused on either extremely narrow solutions (i.e. can simply increasing a driver's pay make them change their mind about quitting?) or esoteric ones, relying on more abstract and subjective features such as a driver's job satisfaction. Our model addresses concerns we have with both of those approaches. Namely, we use a wide variety of feature variables and built the model with the expectation that there would be no one single dominant factor (indeed when we found one in years-at-company, we built a model without using it to try and find deeper causes of turnover). In addition, we focus on hard objective data; something that companies have readily available to them and cannot be misled by human error (except in data entry).

We trained and tested dozens of models with slight variations to find the one that best predicted the data we were given. We discovered dozens of different interactions between feature variables, reinforcing our hypothesis that no single variable is the answer to the problem of driver turnover. We found evidence to support the claim of [CRDG18] that payment variance is at least as much a factor as the wages themselves.

We believe that our work is of great value, particularly to future research. We started with a simple Random Forest model and tuned it to predict driver turnover much better than chance. A more sophisticated model based on our research might do even better still. From a business perspective, we uncovered several variables that are important factors in almost every case, all of which apply to every trucking company and many of which apply outside the trucking industry as well.

In Section 1.1, we stated the goal of our thesis with two questions. This thesis, especially Chapters 4-5, provided answers to them.

> **Q1**: What combination of feature variables, hyperparameters, and data manipulation results in the most accurate driver prediction model?
>
> **A1**: We found a combination of hyperparameters (Section 4.4.1), data views (Table 5.1), and feature variables (Section 5.7.1), that yielded training Evaluation scores as high as 90% and validation Evaluation scores as high as 81%.
>
> **Q2**: Can a model trained on one type of driver successfully predict the behaviours of a different type of driver?
>
> **A2**: Yes, it can. By using a slightly different set of model parameters (Table 5.7), we produced a model even more accurate than the one that

was trained and tested on the same type of driver.

The improvement across different classes of driver suggests to us that the model's only major weakness is a lack of data, and a more robust dataset, even one with different types of drivers, would only strengthen its predictive power.

6.2 Future Work

While we believe we have satisfactorily answered the questions we set to address, there is a great deal more that could be done with this work in the future. Most simply, there are other variations of our base model that we could test. Time constraints limited us in our ability to tweak our model variables - Window Size, for example, was either 90 days or infinite - and there may be better models out there that a more fine-grained search would discover. Other tweaks to the model, such as a time-fading approach to the data, could also be made.

Data was another limiting factor. As discussed in Chapter 5, we had only 80 owner-operator drivers to work with in the validation step, which may have skewed some results. With more data, we could be more confident in our results, and perhaps uncover more relationships between feature variables. Also, our data came from a single company and it would be worthwhile to see if our trained model can successfully predict drivers from an entirely new company.

Our results are interesting from a research standpoint, but from a business perspective, they need some translation. With more time, it should be possible to create an application or a simple process that a company could use to determine if driver X

was likely to turn over in the next year, allowing them to devote resources towards trying to stop that from happening.

In our conclusion, we referenced the fact that many existing studied focus on esoteric reasons for turnover. While we believe in the importance of hard data, we would also like to be able to factor in the opinions of drivers. After all, quitting is a subjective decision, and is likely influenced at least in part by subjective measures our data cannot provide.

Bibliography

[ABS11] Khadija Al Arkoubi, James W. Bishop, and Dow Scott. An investigation of the determinants of turnover intention among drivers. *Loyola University, Chicago*, 5(2):470–480, 2011.

[ABV10] David G. Allen, Phillip C. Bryant, and James M. Vardaman. Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of management Perspectives*, 24(2):48–64, 2010.

[AGZ09] Rajshree Agarwal, Martin Ganco, and Rosemarie H. Ziedonis. Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility. *Strategic Management Journal*, 30(13):1349–1374, 2009.

[AHVM14] David G. Allen, Julie I. Hancock, James M. Vardaman, and D'lisa N. Mckee. Analytical mindsets in turnover research. *Journal of Organizational Behavior*, 35(S1):S61–S86, 2014.

[All19] Canadian Trucking Alliance. Increased truck driver pay lowers turnover rates in 2018, 2019.

[BC55] Arthur H. Brayfield and Walter H. Crockett. Employee attitudes and employee performance. *Psychological bulletin*, 52(5):396, 1955.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[Bue64] William D. Buel. Voluntary female clerical turnover: The concurrent and predictive validity of a weighted application blank. *Journal of Applied Psychology*, 48(3):180, 1964.

[Cas76] Wayne F. Cascio. Turnover, biographical data, and fair employment practice. *Journal of Applied Psychology*, 61(5):576, 1976.

[CRDG18] Samantha Conroy, Dorothea Roumpi, John Delery, and Nina Gupta. Individual pay variability over time and turnover outcomes in the trucking industry. *Academy of Management Proceedings*, 2018:11371, 04 2018.

[DKP81] Dan R. Dalton, David M. Krackhardt, and Lyman W. Porter. Functional turnover: An empirical assessment. *Journal of applied psychology*, 66(6):716, 1981.

[Faw06] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[FFL76] Suzanne M. Federico, Pat-Anthony Federico, and Gerald W. Lundquist. Predicting women's turnover as a function of extent of met salary expectations and biodemographic data. *Personnel Psychology*, 1976.

[FH62] Edwin A. Fleishman and Edwin F. Harris. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 15(1):43–56, 1962.

[FMH+09] Will Felps, Terence R Mitchell, David R Hekman, Thomas W Lee, Brooks C Holtom, and Wendy S Harman. Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of Management Journal*, 52(3):545–561, 2009.

[FPS+96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy, et al. *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park, 1996.

[Fri01] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[HB17] Mitchell Hoffman and Stephen V. Burks. Worker overconfidence: Field evidence and implications for employee turnover and returns from training. Technical report, National Bureau of Economic Research, 2017.

[HC18] David Hand and Peter Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, 2018.

[HLSH17] Peter Hom, Thomas Lee, Jason Shaw, and John P. Hausknecht. One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102, 01 2017.

[Hom11] Peter W. Hom. Organizational exit. *APA handbook of industrial and organizational psychology*, 2:325–375, 2011.

[Hul66] Charles L. Hulin. Job satisfaction and turnover in a female clerical population. *Journal of Applied Psychology*, 50(4):280, 1966.

[HW73] Don Hellriegel and Gary E. White. Turnover of professionals in public accounting: A comparative analysis1. *Personnel Psychology*, 26(2):239–249, 1973.

[KN73] H.B. Karp and Jack W. Nickson. Motivator-hygiene deprivation as a predictor of job turnover. *Personnel Psychology*, 1973.

[LDP19] Yingjie Lu, Shasha Deng, and Taotao Pan. Does usage of enterprise social media affect employee turnover? empirical evidence from Chinese companies. *Internet Research*, 29, 03 2019.

[LJR08] Jorge M. Lobo, Alberto Jimnez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008.

[LL17] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[LM94] Thomas W. Lee and Terence R. Mitchell. An alternative approach: The unfolding model of voluntary employee turnover. *Academy of management review*, 19(1):51–89, 1994.

[LT89] Stephen A. LeMay and Stephen Taylor. The truck driver shortage: An overview and some recommendations. *Journal of Transportation Management*, 1(1):47–55, 1989.

[McN18] Sean McNally. Turnover rate at large truckload carriers rises in first quarter, 2018.

[MGHM79] William H. Mobley, Rodger W. Griffeth, Herbert H. Hand, and Bruce M. Meglino. Review and conceptual analysis of the employee turnover process. *Psychological bulletin*, 86(3):493, 1979.

[MGS+19] Tiago Massoni, Nilton Ginani, Wallison Silva, Zeus Barros, and Georgia Moura. Relating voluntary turnover with job characteristics, satisfaction and work exhaustion - an initial study with Brazilian developers. *CoRR*, abs/1901.11499, 2019.

[MH18] Emma McIntosh and Trevor Howell. It's just a step in the process. families relieved as charges laid against calgary truck driver in humboldt broncos bus crash. *The Star*, 2018.

[Mob77] William H. Mobley. Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of applied psychology*, 62(2):237, 1977.

[Pav16] Christina Pavlou. How to calculate employee turnover rate. 2016.

[PDP19] PDPBox. PDPBox documentation, 2019.

[Pow07] D. Powers. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation (tech. rep.). *Adelaide, Australia*, 2007.

[Pow15] David M.W. Powers. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. *arXiv preprint arXiv:1503.06410*, 2015.

[PS73] Lyman W. Porter and Richard M. Steers. Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological bulletin*, 80(2):151, 1973.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RLTT94] Michael D. Richard, Stephen A. LeMay, G. Stephen Taylor, and Gregory B. Turner. An investigation of the determinants of extrinsic job satisfaction among drivers. *The International Journal of Logistics Management*, 5(2):95–106, 1994.

[Sch67] Allen J. Schuh. The predictability of employee tenure: A review of the literature. *Personnel Psychology*, 1967.

[SKH69] P.C Smith, L.M. Kendall, and C.L. Hulin. *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes.* ERIC, 1969.

[SO74] Donald P. Schwab and Richard L. Oliver. Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology*, 27(1):125–128, 1974.

[SOK19] Koya Sato, Mizuki Oka, and Kazuhiko Kato. *Early Turnover Prediction of New Restaurant Employees from Their Attendance Records and Attributes*, pages 277–286. 08 2019.

[Ste02] Robert P. Steel. Turnover theory at the empirical interface: Problems of fit and function. *Academy of Management Review*, 27(3):346–360, 2002.

[US 19] US Department of Labor. Bureau of labor statistics, 2019.

[Wel18] Megan Wells. Employee turnover rates in the healthcare industry, 2018.

[ZMXX19] Xin Zhang, Liang Ma, Bo Xu, and Feng Xu. How social media usage affects employees job satisfaction and turnover intention: An empirical study in China. *Information & Management*, 56(6):103–136, 2019.