POST-MIDTERM PROJECT: WORKERS' COMPENSATION CLAIMS MODELING

Ngaitik Chung, Moncef El Ouriaghli, Xinxin Lai, Sachin Varule, and Ryan von Trapp

University of North Carolina – Charlotte

Table of Contents

Table of Figures

Tables

Figures

Part I: Executive Summary

Post-Midterm Project Objectives

The objectives for the post-midterm project are to use the data prepared in the mid-term project submission to (1) visualize and explore the data in Tableau, (2) create a logistic regression model, and (3) provide strategic recommendations for improving Workers' Compensation claims processing.

Initial Hypothesis and Key Findings

Initial hypothesis while reviewing the data were that the injury type and income level of the claimant may be related to the overall cost of the Workers' Compensation claim. Time also appeared to be a factor, but neither total cost nor total time were aggregated values in the data.

Three calculated factors were added to the dataset to measure total payout, total claim duration (injury to case closing), and number of days from the injury to filing the claim. Key findings after visualization were that three factors contributed to over half the total cost of Workers' Compensation claims, and that litigation increased the cost of claims by 10 to 20 times the amount of non-litigated claims. A logistical regression model was then built to predict whether the claim will be litigated.

Recommendations

Two key recommendations are made from this project related to cost control. First, focus on reducing the top three injuries that result in the highest total cost: strains, contusions, and sprains. Second, use machine learning to predict if a claim is likely to become litigated and proactively work with those claimants to resolve the claim outside of litigation.

Part II: Post-Midterm Report

Introduction

This report details the expected relationships between each of the independent variables, findings from data visualizations, the logistic regression model, and strategic recommendations for the firm.

Expected Relationships Between Independent Variables

Initial analysis by this group into Worker's Compensation claims processes focused on the total cost of claims in relation to service duration and injury type. Early research performed in preparation for the mid-term report indicated that claim cost has a positive correlation to the time between injury and claim filing by the claimant, and the total duration of a claim (injury date to case closing). Other factors that were initially explored were related to fatality rate, worker age, and salary. The thought was high-cost injuries, such as lost limbs or fatalities, would be the most expense claims on which to focus attention. Worker age was thought to be a factor, under the assumption that younger workers may be more reckless, or that older workers are more prone to injuries and disease. Salary is considered as a contributing factor, since lost wages are a cost included in Workers' Compensation claims.

In order to further explore these assumptions, aggregate data was needed for total payment amount, total duration of the claim, and the time between injury and claim filing. Three aggregated factors were added to the dataset to address this need:

- SUM_of_PaymentAmount: Sum of total payments.
- ServiceDuration: The duration of the claimant service (ClaimClosedDate - ClaimOpenedDate).

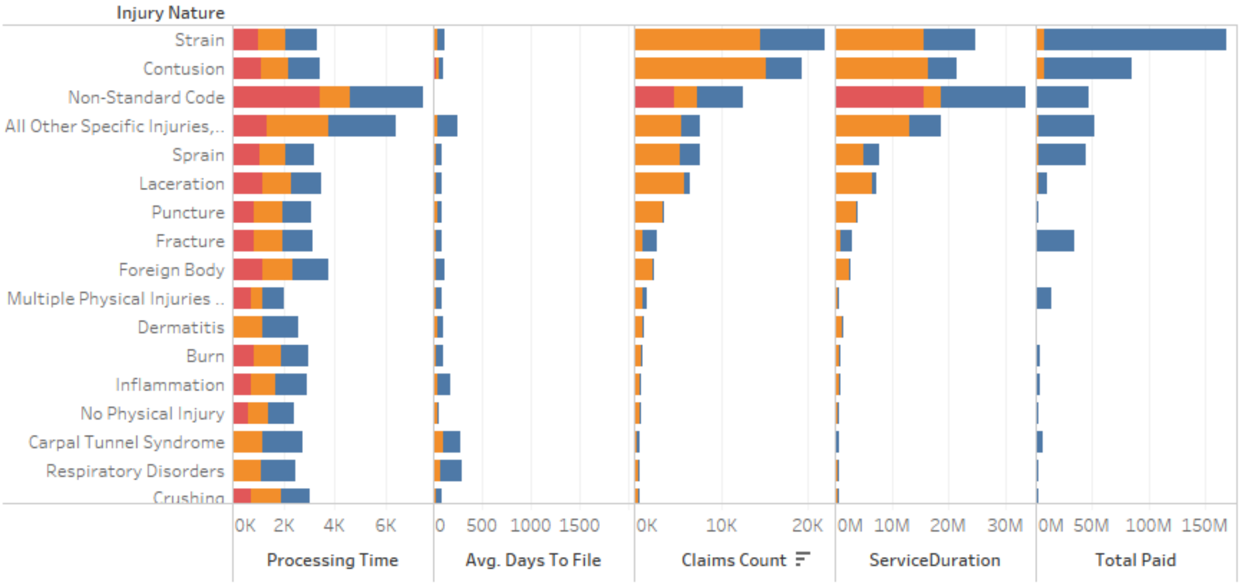
- DaysToFile: The duration between the incident date and claimant opened date (ClaimOpenedDate - IncidentDate).

Next, the data were imported into Tableau Desktop for visual exploration.

Data Visualizations in Tableau Desktop

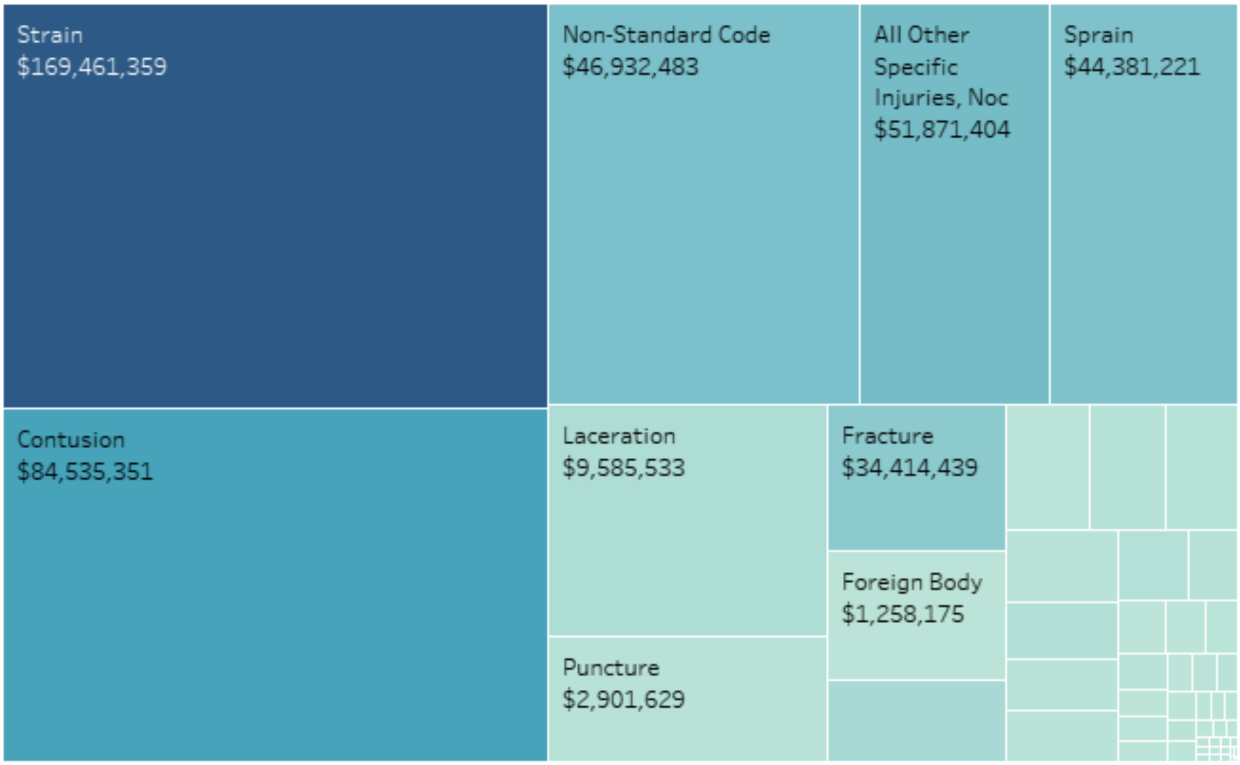
The data were initially visualized by comparing each of the factors against total and average claims costs. After several iterations, a pattern imerged where the highest total claims cost appeared related to a handful of the most-commonly occuring injury types: strains, contusions, and sprains. (Other non-specific codes, "Non-Standard Code" and "All Other Specific Injuries" occurred frequently as well, but were excluded from analysis due to a lack of specificity in injury type.) These three injury types also occurred most frequently, and totaled the highest service duration:

Figure 1. Summary of Claims Measures



To better understand the correlation between claims frequency, claims count, and cost, a tree map was created to compare the total count of each injury type to the frequency of claims filed. Figure 2 illustrates this relationship:

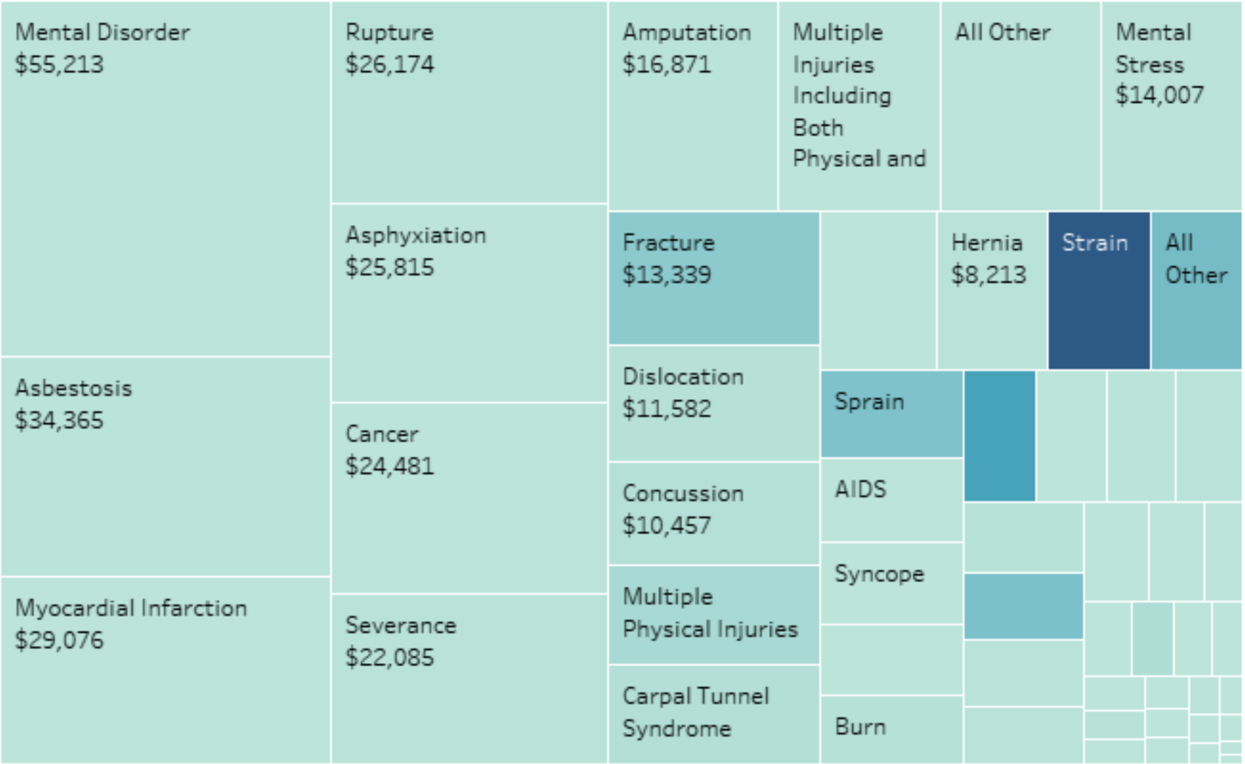
Figure 2. Count of Injury Type and Total Cost



The size of each box correlates to the total count of claims. Dark-blue shading indicates the highest total cost, while light green is lowest total cost. Strains, contusions, and sprains accounted for 52% of the claims filed, and 59% total cost for claims. While other non-specific claim types (e.g., "Non-Standard Code") accounted for a large proportion of claims and cost, the goal of this report is to provide actionable recommendations, which require the specificity in the injury type.

While the impact of total claim cost was evident, the group also wanted to understand the average cost for strains, sprains, and contusions.

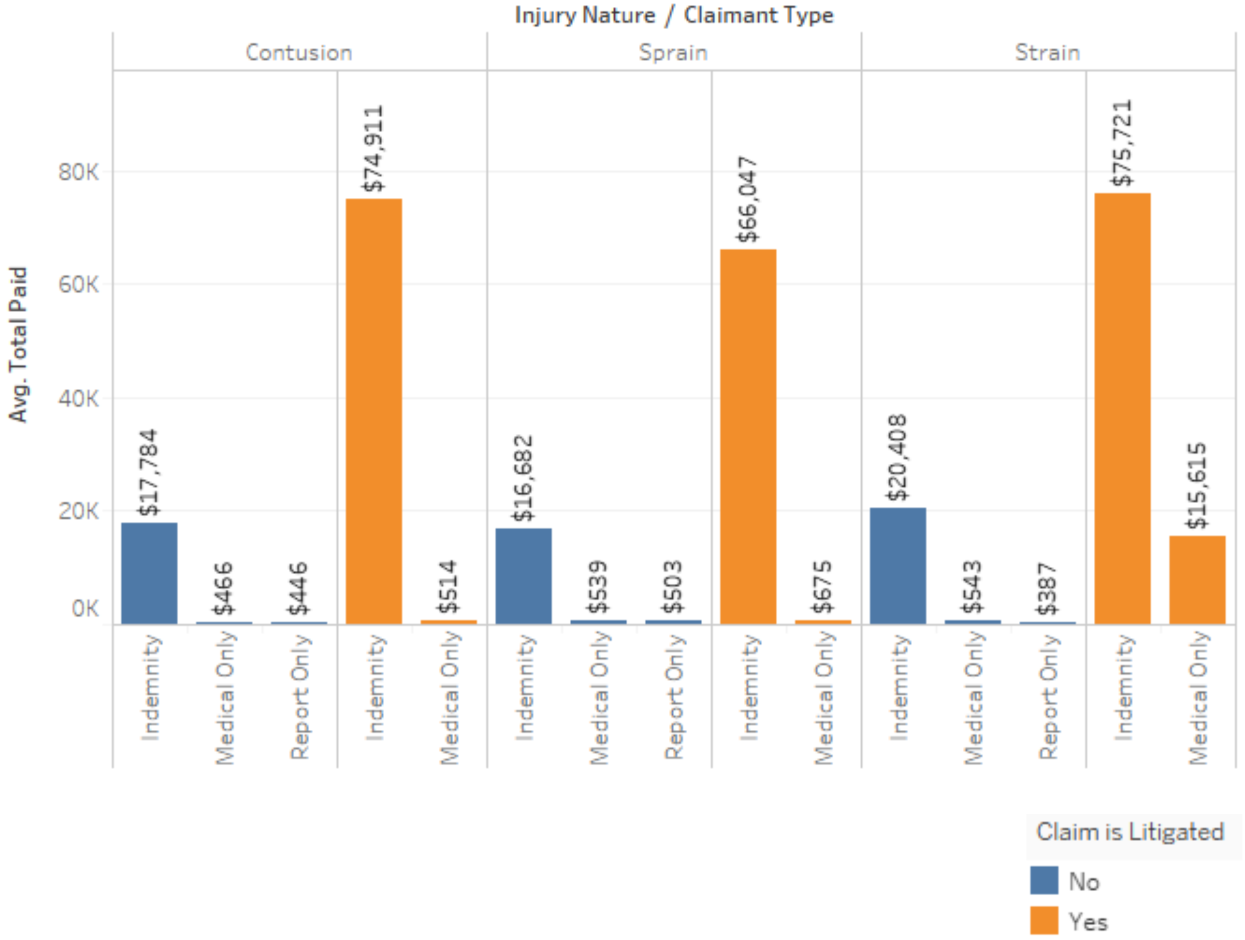
Figure 3. Average Cost Per Claim



Again, the color of the boxes indicates the total cost (dark blue is high total cost), but the size of each box is average claim cost. Strains, sprains, and contusion average cost is slightly higher than average. A key finding from this analysis is that these three injury types, while essentially average in cost per claim, accounted for the highest proportion of total cost due to the high number of claims filed. Thus, one way of reducing total cost is to reduce the number of these claims by introducing programs to mitigate the risk of strain, sprain, and contusion injuries.

Other costs associated with these claims were explored, and a notable increase in claims cost was discovered for litigated versus non-litigated strain, sprain, and contusion injuries, as shown in Figure 4.

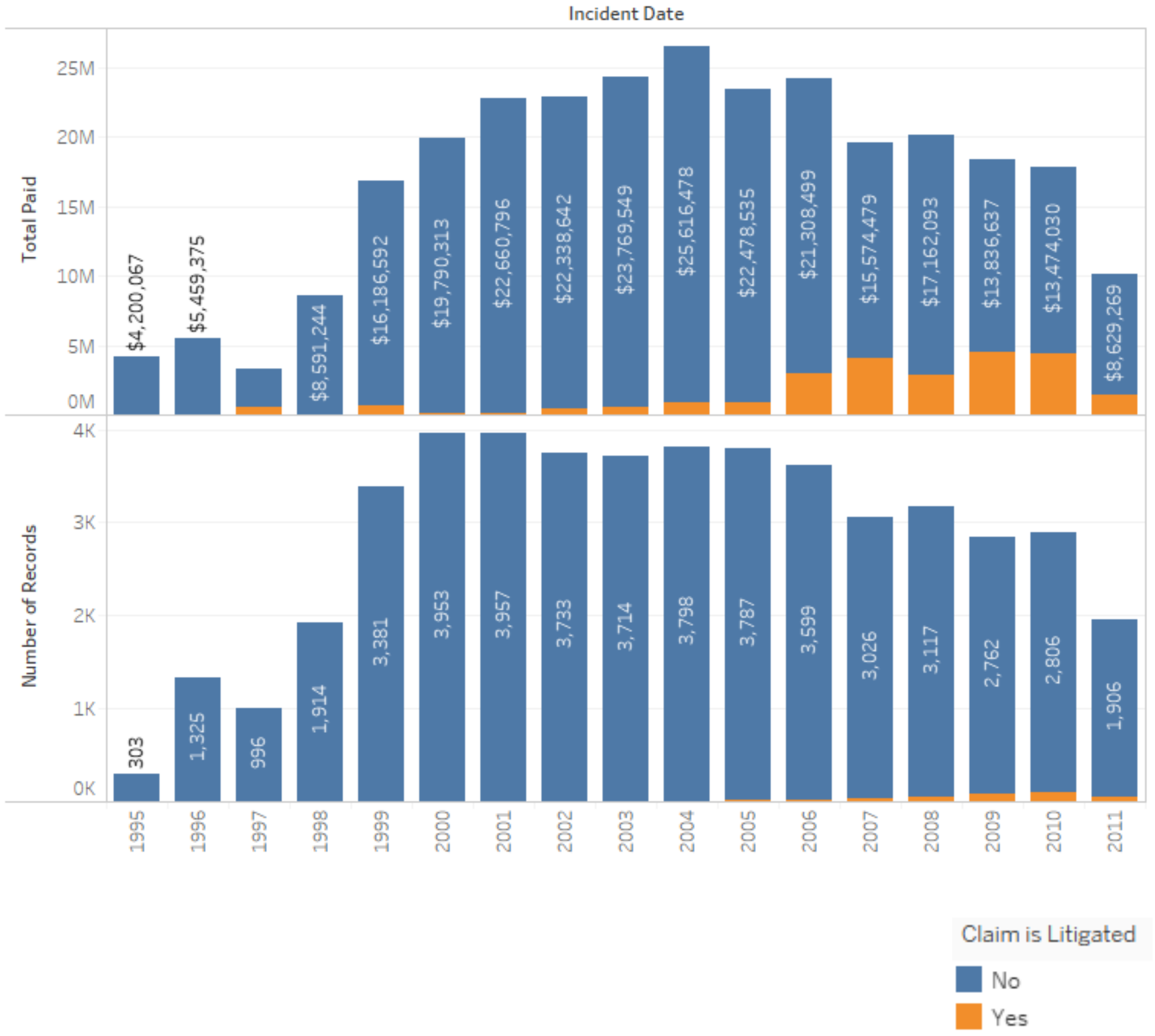
Figure 4. Average Cost of Litigated vs. Non-Litigated Claims



Overall, the average cost of a litigated claim was about five times the cost of a non-litigated claim. This increase in cost was most notable with indemnity claims. While the underlying reason for indemnity is unknown, a generalization could certainly be made that litigation has a significant increase on the average cost per claim. This pattern was also true for other injuries, with the average increase in litigated claims between 10 to 20 times the non-litigated claim cost.

A pattern of increased litigation was also discovered for strains, sprains, and contusions since 1995, as illustrated in Figure 5:

Figure 5. Strain, Sprain, and Contusion Claims Since 1995



Thus, the second key finding is that litigation significantly increases individual claims cost. Developing a logistic regression model to predicts whether the claim may become litigated can potentially reduce the number of litigated claims by enabling the firm to proactively identify claimants who require case management. The concept is that the claimant will still receive the care and financial resources needed, but legal costs can be eliminated.

Logistic Regression Model

R Studio was used to develop a logistic regression model with an outcome variable of "IsLitigated". In short, the regression model predicts the likelihood that a claim will become litigated. All logical decision variables were included in the initial development of the model, but the number, and variety, of predictors caused issues with the predictive capabilities of the model. The number of predictor variables was eventually reduced to the following through trial and error:

- IsFatality
- ClaimantAge_at_DOI_Computed
- Gender
- ClaimantType
- InjuryNature
- DaysToFile

InjuryNature originally utilized all categorical values for injury. However, that resulted in an error where the outcome variable was predicted to be either 0 or 1 (instead of a continuous value in between 0 and 1). As a result, this variable was grouped into three categories: strain, non-standard code, and contusion (an assumption was made that strains and sprains were similar in nature regarding litigation).

The resulting logistical model is summarized in the following output from R Studio:

```
Call:
glm(formula = IsLitigated ~ IsFatality + ClaimantAge_at_DOI_Computed +
    Gender + ClaimantType + InjuryNature + DaysToFile, family = binomial(link
= "logit"),
    data = df.train.Combined)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.87324  -0.33194  -0.27856  -0.00015   2.56799

Coefficients:
```

```
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -1.129e+00  4.867e-01  -2.319 0.020375 *  
IsFatality1                -2.713e+00  4.353e-01  -6.232  4.6e-10 ***
ClaimantAge_at_DOI_Computed  1.388e-02  9.025e-03   1.538 0.123996    
GenderMale                  1.254e-01  1.824e-01   0.688 0.491759    
GenderNot Available        -9.768e-01  8.220e-01  -1.188 0.234701    
ClaimantTypeMedical Only   -3.379e+00  2.384e-01 -14.173  < 2e-16 ***
ClaimantTypeReport Only    -1.485e+01  5.092e+02  -0.029 0.976733    
InjuryNatureOther           8.153e-01  2.599e-01   3.137 0.001706 ** 
InjuryNatureStrain          9.586e-01  2.706e-01   3.542 0.000397 ***
DaysToFile                  2.731e-04  2.678e-04   1.020 0.307825    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1269.10  on 1193  degrees of freedom
Residual deviance:  787.78  on 1184  degrees of freedom
AIC: 807.78

Number of Fisher Scoring iterations: 16
```

The model was tested using a test sample of approximately 92,000 rows, with a probability predictor cutoff of 0.3. Overall performance of the model was satisfactory, with 81.8% accuracy, 82.0% True Positive Rate, 18.2% False Positive Rate, and an AUC of 85%, as shown in the following tables and figures:

Table 1. Confusion Matrix

Predictors	**Actual**		**Total**
	Not Litigate	**Litigate**	
Not Litigate	75,111 (TN)	48 (FN)	75,159
Litigate	16,697 (FP)	219 (TP)	16,916
Total	91,808	267	92,075

Accuracy

$$\frac{TP+TN}{All\ Cases} = \frac{75{,}111 + 219}{92{,}075} = 81.8\%$$

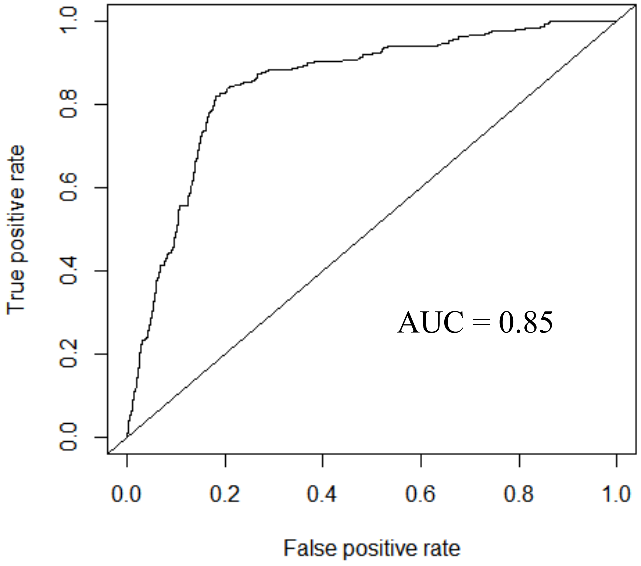
True Positive Rate

$$\frac{TP}{TP+FN}=\frac{219}{218+48}=82.0\%$$

False Positive Rate

$$\frac{FP}{FP+TN}=\frac{16{,}697}{16{,}697+75{,}111}=18.2\%$$

Figure 6. ROC Curve



In general, this model is a good predictor of which cases will likely be litigated and can be used to efficiently determine which Workers' Compensation cases require management.

Strategic Recommendations

Two recommendations are made for the firm to lower the cost of Workers Compensation claims, based on insights into the most common claims.

Reduce Strain, Sprain, and Contusion Injuries

First, focus attention on minimizing the number of sprain, strain, and contusion injuries. While the "Non-Standard Code" and "All Other Specific Injuries, Noc" categories also incurred high costs, these categories lack the specificity needed to direct resources towards the root cause of injuries. With support from leadership at the firm, corporate resources can be directed towards better understanding the underlying cause of sprain, strain, and contusion injuries to develop policies and procedures that will reduce injuries. Safety and training programs can also be improved from insights gained from root cause analysis. These three categories represent 52% of all claims filed and the highest total cost to the firm and the number of claims filed has significantly increased since 1995. To justify the cost of program development, it should be noted that each 1% reduction in the number of sprain, strain, and contusion injury claims equates to approximately $250,000 in annual cost savings.

Use Machine Learning to Predict Litigation Cases

Currently, the firm does not utilize predictive analytics to determine if a case will likely be litigated. All cases are assumed to have equal probability, and it is likely that case managers use heuristics to determine which cases require management. As described earlier in this report, litigated claims are approximately 10 to 20 times more costly than non-litigated claims. Using machine learning to predict which cases may go to litigation offers an opportunity for lowering Workers Compensation claims cost without adversely affecting a positive outcome for the claimant.

To further illustrate the average cost of litigated claims, approximately .05% of all claims went to litigation, but represented approximately 7% of total claims cost and were over 10 times the average cost of non-litigated claims.

Table 2. Total Litigated vs. Not Litigated Claims

	Litigated	**Not Litigated**
Count of Claims	534	92,735
Total Payout	$34,388,779	$469,920,351
Average Payout	$64,398	$5,067

Using basic information about the claim (e.g., injury nature, age of claimant, gender, etc.), the firm can predict if the claim is likely to go to litigation. Claims that flag as litigation cases can then be assigned a case manager and proactively managed to reduce the likelihood of litigation.