**WORKERS COMPENSATION CLAIMS PROCESSING**

**FINAL APPENDIX**

**Prepared By**

*Shashank Lakshminarayan Suresh*

*Noelline Dsouza*

*Vinaya Tinguria*

*Esra Alshamari*

*Jesudamilare Ibitoye*

*Aysha Nahan Mohiyeddeen*

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SECTION #** | **SECTION** | **PAGE NUMBER** |
|  | **EXECUTIVE SUMMARY** | **3** |
|  | **OVERVIEW** | **4** |
| **1.** | **DATA MERGE** | **4** |
| 1.1 | INDEPENDENT DERIVED VARIABLES | **4** |
| 1.2 | DEPENDENT DERIVED VARIABLES | **4** |
| **2.** | **DATA VISUALIZATION** | **5** |
| **3.** | **PREDICTIVE MODELING** | **8** |
| 3.1 | OVERVIEW OF COMPARISON OF THE MODELS | **8** |
| 3.2 | LOGISTIC REGRESSION | **8** |
| 3.3 | DECISION TREE | **9** |
| 3.4 | NAÏVE BAYES CLASSIFIER | **10** |
| 3.5 | LINEAR REGRESSION | **11** |
| 3.6 | SUMMARY OF COMPARISON | **12** |
| 3.7 | SIMPLE IMPLEMENTED MODEL FOR COMPARISON OF VARIOUS TECHNIQUES | **13** |
| **4.** | **BUILDING PREDICTIVE MODEL** | **14** |
| **5.** | **INSIGHTS** | **20** |
| 5.1 | INSIGHTS FROM THE VISUALS | **20** |
| 5.2 | INSIGHTS FROM THE PREDICTIVE MODELLING | **21** |
| **6.** | **RECOMMENDATIONS** | **21** |
| **7.** | **ANALYTICS 3.0** | **22** |

**EXECUTIVE SUMMARY:**

The goal of this analysis is to advice a company who works with workers compensation rights on how to use data science to help identify drivers of cost and process time. Information concerning background on workers’ compensation rights and data exploration of the claims data were completed earlier and presented. This report will focus primarily on the modeling and recommendations for the company.

One key insight that was made according to fig. 2.1 is the fact that the most incidents were reported on Wednesday and the least incident occurred during the weekend. This takes away our hypothesis of the most incident occurring on Monday. We thought fraud cases could be found if this was true because people might be getting hurt outside of work; then coming to work and filing a claim. We recommend further investigation and analysis should be made to determine if people are truly getting injured on Wednesday or are getting injured outside work and just waiting to file it.

Another insight is that the maximum off days for workers reported that they had mental disorders and psychological problems (fig 2.3). Based on this, we would recommend that the company offer counseling to the employees so that they could go in for a session. Then, the doctor could then diagnose it and figure out the source of the problem. This would help save the company money in the long run if they can determine the source of the problem and prevent it.

According to figure 2.4, we noticed that the average process time remained until 2011. From the year 2011, there was an increase in the average process time. One could assume that the process that they reviewed the cases was change, but one cannot go on assumptions without having the information to back it up. From our predictive model, decision tree, we were able to see that thirty two percent of indemnity claims are high risk. We could also determine that the injury nature has an impact on the process time. We combined the observations, as well as the predictive model to recommend that one should assign cases based on whether or not it is a high risk. This should be done by first determining if a risk is high, low, or medium based on the decision tree. Then, give cases with high risk to people with more experience so that overall the process time will reduce.

All in all, the use of analytics would help save money for the company and would effectively help in determining fraud. In order for the process of analyzing the data to be smoother, they would need to be a chief analytics officer. They would be the one to lead the operation and help in monitoring how analytics could help the company. Also, implementing agile analytics methods would help in increasing the speed. Analytics is becoming more popular in helping to predict how the company does in the future. By getting ahead of the game, the company gets a competitive advantage over other company.

**OVERVIEW:**

The below documentation explains the steps that leads to a predictive model based on which the Analytics team can advise the Workers Compensation Claims company on reducing the main cost drivers the company currently faces. In the first part, the team explored the data and arrived at a cleaned data set. In the following steps, we focused on merging the claims dataset with transaction data, adding the derived variables, exploration of the merged data, comparison of possible predictive models, building the predictive models and in giving strategic recommendations for the Workers Compensation Claims Processing Company.

**1. DATA MERGE:**

We merged cleaned claims data with transactions data based on the Claims ID column. We used R to do this task. We imported both claims and transactions data set into R and merged them both. Before the merge, our cleaned Claims data had 133971 records. After the merge, the merged dataset had 108617 records. The difference in the record count is due to the fact that the transactions data had less records and since the merge was based on Claim ID, the Claim ID’s that were not present in transactions dataset were lost.

**1.1** Independent Derived Variables:

We selected two independent derived variables. The first one is the ‘Off Days’. ‘Off Days’ represents the number of days the employee did not come to the office due to the injury. Off Days is calculated by taking the difference between Incident Start Date and Return to Work Date. The reason we selected Off days as one of the independent derived variables is because it gives us an indication as to what the employee had to endure. It also helped us to identify the severity of the claim and would be useful in estimating the cost.

The second independent derived variable is ‘Day of Incident’. This gave us the weekday the incident occurred (like Sunday, Wednesday, etc..). We can obtain the value for Day of Incident by looking at the Incident Date column. This column helped us to find out if there was any relationship between the Day of Incident and the average claims per day. Using this variable, we can identify if there is a specific day where the claims tend to be significantly higher or lower than the rest. This data is also useful to estimate if the employee is claiming the worker’s compensation on Monday when the incident actually happened on Saturday or Sunday outside work.

**1.2** Dependent Derived Variable: (High-risk/Low-risk)

In order to classify the claims as High or Low risk, we need to identify the variables which we can be used to make this classification. We used Total Paid, Off Days and Process Time as the three variables which would decide whether the claim is a high-risk or a low-risk claim. A cut-off condition was defined for each of these variables and if the claim met any one of those conditions, then it was classified as a high-risk claim. The cut-off numbers used were calculated by taking the average of their respective columns and multiplying it by 1.5 magnitudes.

The formula used for classification was:

IF (Total Paid>12000 or Off Days>70 or Process Time>2400) => High Risk, Else => Low Risk

After the classification, we saw that around 30% of the claims were high risk claims and rest of them was low risk claims.

**2. DATA VISUALIZATION**

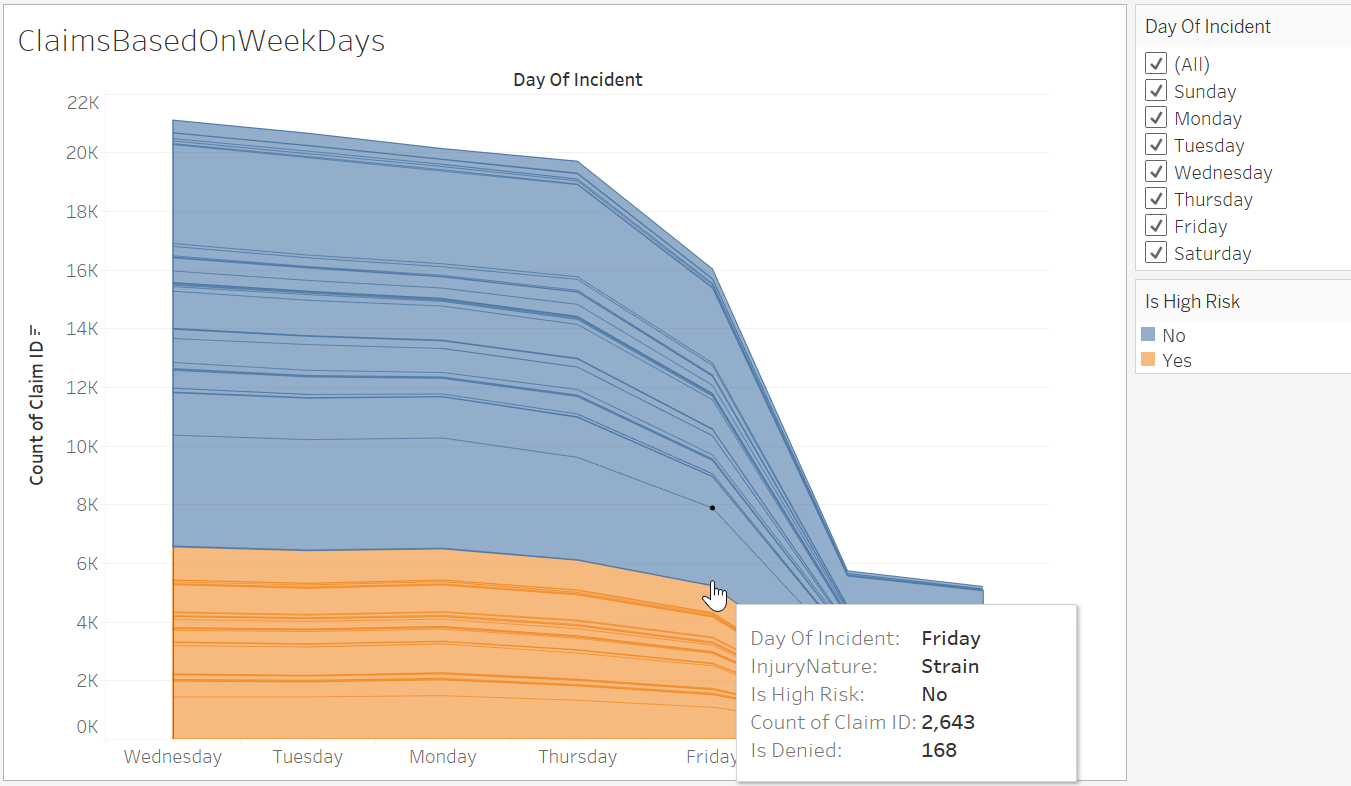
Both claims and transactions datasets were merged and cleaned to have a more holistic overview of the claims processing process. To have even further understanding and analysis, we derived three variables out of the existing dataset. The additional variables following merging of the datasets are as below:

* **Off Days:** Calculated by subtracting day of incident from date back to work.
* **Day of Injury:** links each date of injury with the respective day of the week
* **Is High Risk:** derived by deciding on specific thresholds of claims processing time and cost.

We have further analyzed the dataset in correspondence to the derived variables. The additional insights gained by adding the newly derived variables are below.

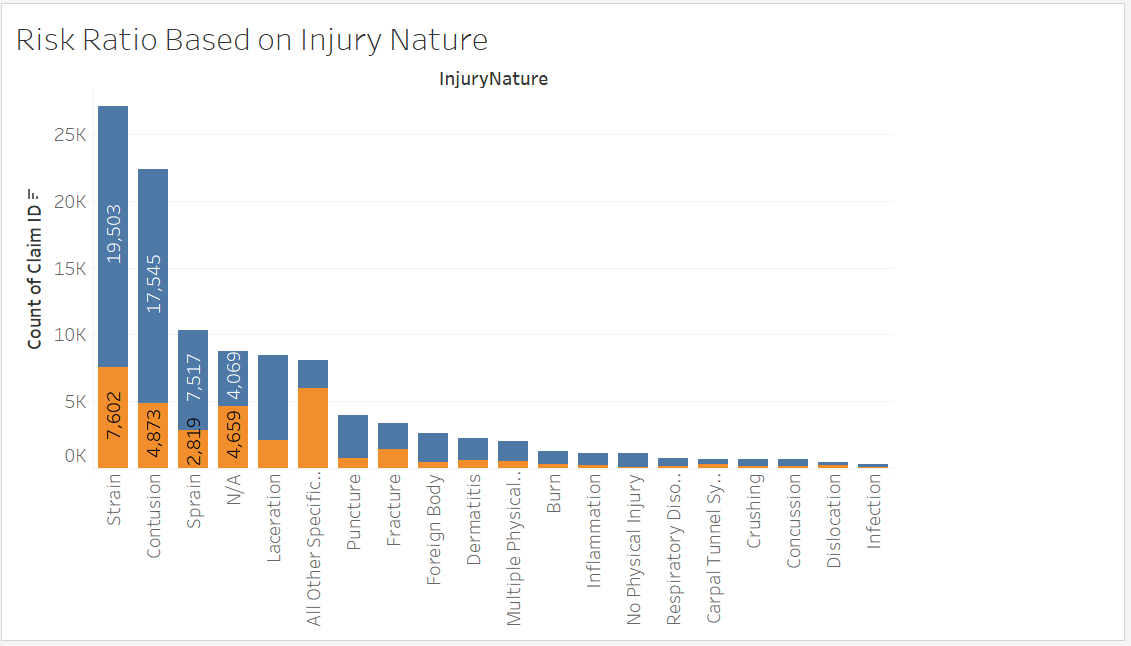
Below are the visualizations results. Each of these visualizations provides additional insight of the datasets that provides more evident and strong base for enhancement plan of the claims processing by the insurance company.

**Fig 2.1**

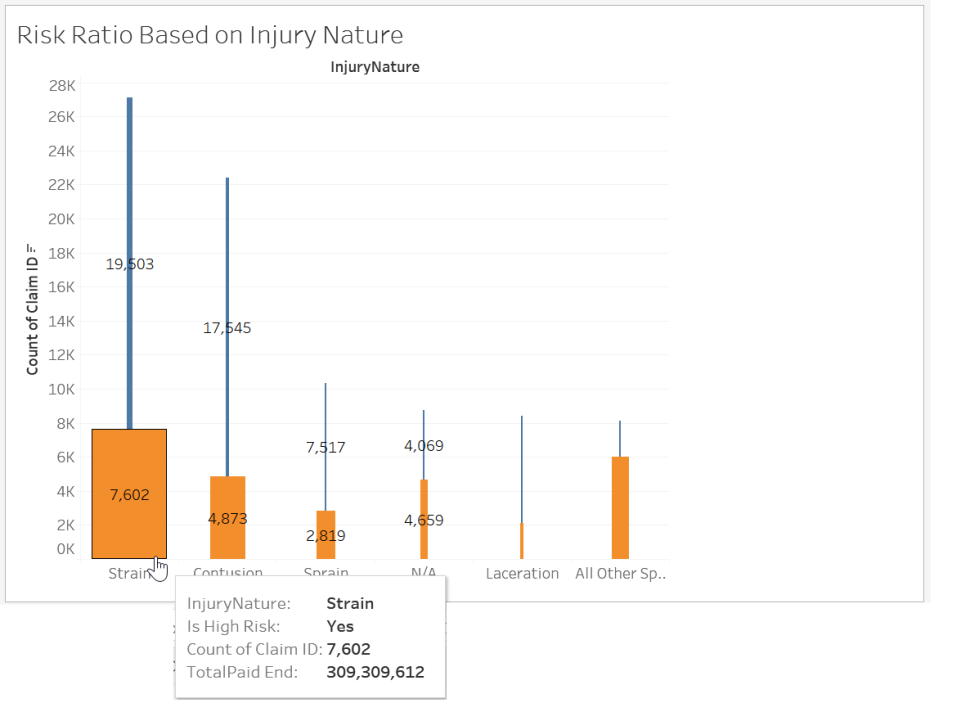


The barchart shown provides count of claims along with ratio of high-risk claims categorized into different injury natures. The highest occuring injury nature are strain, contusion and sprain. Next to these categories is the N/A category which need to be further governed/properly classified by the insurance company.

**Fig 2.2**



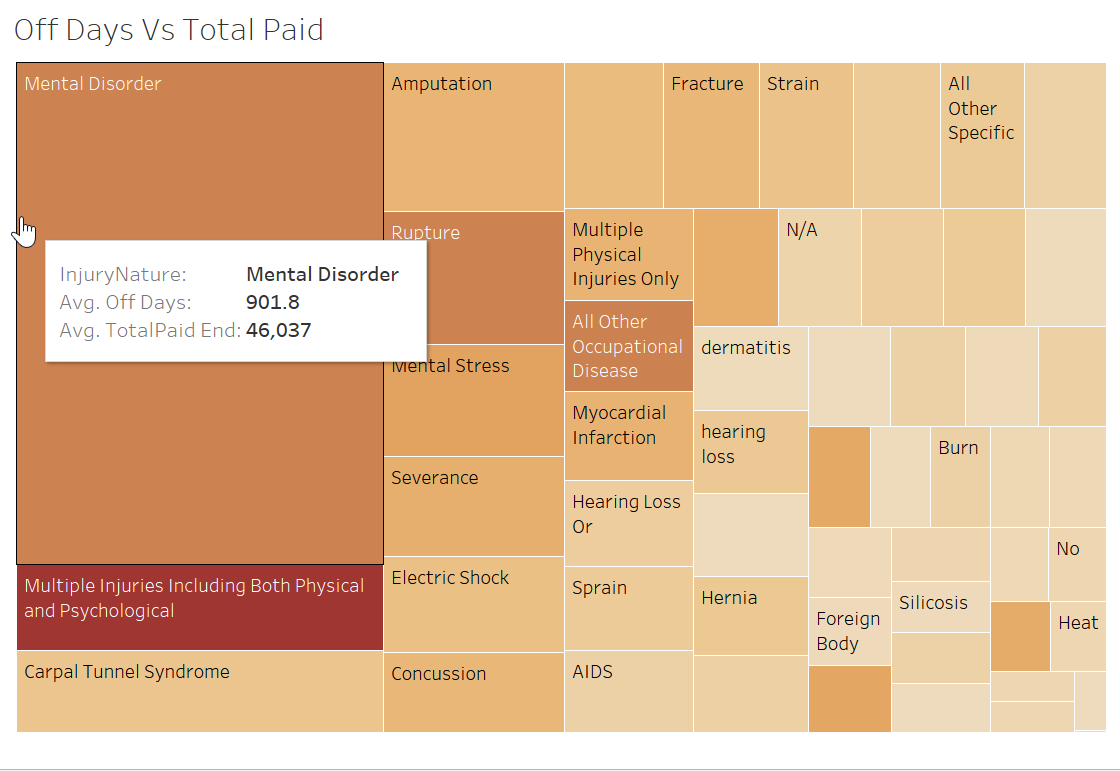
The injury nature categories with highest ratio of High risk claims is the ‘All other specifics’. 75% of the named category are tagged with high risk. Further drilling down, we can see that the high risk is nearly with equal causes of having long processing time and high total paid amount. In addition, the ‘N/A’ category is having almost 50% high risk claims which again routes back to the point that insurance companies should have revised categories and strict governance on injury nature categories.

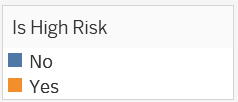


**Fig. 2.1.1**

Further drilling down to categories with less count of claims, we can see that the ratio of high risk claims drops dramatically. Almost all categories with less than 4K claims count have got a high risk % less than 25%. These categories have got less impact compared to the top 6 categories.

**Fig 2.3**

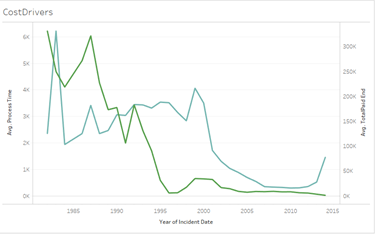


The highest category with offDays is the mental disorder, followed by multiple injuries and Carpal Tunnel Syndrome which is associated as well with high funds paid. Moreover, the highest categories having highest amount paid is the multiple injury category. This category has got 53 cases under it.

It is noticeable that the average total spent for cases with carpal tunnel syndrome is $11K with average 147 offDays. There are 686 cases with CTS out of which nearly 50% cases are having high risk.

Below line graph shows the average processing time and totalPaid amount on a time series. The main cost drivers of claims are not correlated. Both costs were on their peaks during early 80s. That was probably due to lack of digital information and automated processing of claims. The average of both cost drivers have dropped in the 21st century to below than 100K.

**Fig 2.4**



**3. PREDICTIVE MODELING**

**3.1** **Comparison of Models**

The cost drivers for the Workers’ Compensation Claims Processing Company can be broadly divided into two:

1. Amount Paid
2. Processing Time.

As a Data Analytics team, we can give the Claims company some strategic recommendations based on the models that we built. In order to achieve this, we considered four predictive modeling techniques for which the target(dependent) variable and the independent variables are selected from the cleaned dataset. **The four modeling techniques are Linear regression, Logistic regression, Decision tree and Naïve Bayes.**

The possible independent variables to predict are decided based on the questions that we addressed in order to reduce the cost drivers, which were:

* How to reduce the Total expenditure? (**Total\_Paid\_end**)
* How to reduce the Claims Processing time? (**Processing time**: difference between claim open date and closed date)
* How to identify the claims that should be denied? (**IsDenied**)

**3.2 Logistic regression**:

Characteristics of logistic regression that was useful for modeling this dataset:

Logistic regression is a predictive technique used when the dependent variable has binary outcomes. It is considered a “supervised learning technique” since the data includes actual outcome values from past observations. It does not need a linear relationship between the dependent and independent variables. [Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/)can handle all sorts of relationships because it applies a non-linear log transformation to the predicted odds ratio. It can handle ordinal and nominal data as independent variables. Moreover, logistic regression can handle large sample data set; however, the model should be fitted correctly. A good approach to ensure this is to use a stepwise method to estimate the logistic regression so; we used the stepwise method to derive at the independent variables.

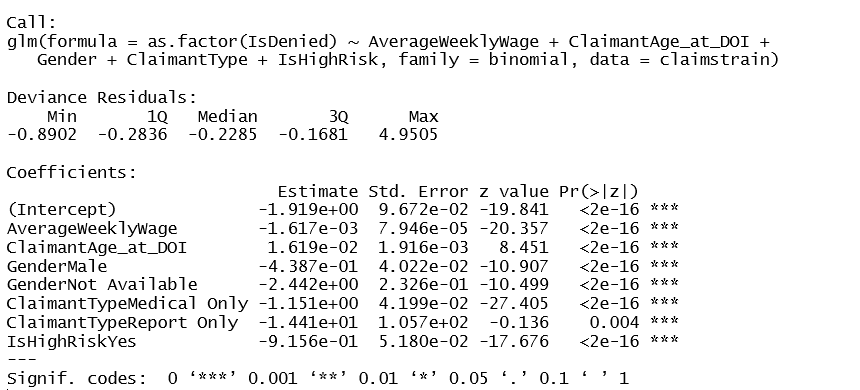
Since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the dependent variable is coded accordingly.

We selected IsDenied as the dependent variable which met this criterion.

Dependent variable: IsDenied

Stepwise regression gave the best model as:

**Fig 3.1**



The advantages of logistic regression come at a cost. In logistic regression, the variables should not be correlated. Also, it requires much more data to achieve stable, meaningful results. With standard regression, typically twenty data points per predictor is considered the lower bound. For logistic regression, at least fifty data points per predictor is necessary to achieve stable results.

Since we have a huge data set, logistic regression is a good candidate modeling technique.

**3.3 Decision tree**:

This is a supervised segmentation technique, which helps to find the trend or pattern in the data by grouping them. If every member of a group has the same value for the target, then the group is pure. If there is at least one member of the group that has a different value for the target variable than the rest of the group, then the group is impure. Ideally, we would like the resulting groups to be as pure as possible i.e. homogeneous with respect to the target variable. Entropy is a measure of disorder that can be applied to a set, such as one of our individual segments, which leads to a further split.

Characteristics that we considered for modelling are:

* They easily handle feature interactions and they’re non-parametric, so you don’t have to worry about outliers
* When fitting a decision tree to a specific dataset, the top few nodes are usually the most important variables within the entire dataset.
* They are also good to solve problems such as scale differences. For example if we had a dataset that measured in average house per state in thousands and average age of people that owned houses in years, those are two units of measurements. If one was to use a regression model, one would have to have some sort of normalization and interpretation of the coefficients. With decision trees, this is not required because the structure of the tree stays the same regardless of the change.
* They are also easy to explain, understand, and tell you about you information.

The main drawback of decision tree is overfitting. An induced tree may over-fit the training data which will result in:

* Too many branches, reflecting anomalies due to noise or outliers
* Poor accuracy for unseen samples

The tree has to be pruned in order to avoid overfitting.

Possible dependent variable

1. **Risk Factor** (High risk/Low risk based on processing time)

Independent variables selected: AverageWeeklyWage, Body\_Part, BodyPartRegion, ClaimantAge\_at\_DOI, ClaimantStatus\_End, ClaimantType, ClaimID, DayOfIncident (Day of the Week), Gender, Injury\_Nature

1. **IsDenied**(Whether the claim is denied or not)

Independent variables:ClaimantStatus\_End + AverageWeeklyWage +

ClaimantAge\_at\_DOI + Gender + ClaimantType + IsHighRisk

**3.4 Naïve Bayes**:

This technique uses Bayes theorem in predicting the posteriori probability of hypothesis using the following basic equation:

p(H|X) = p(X|H) \* p(H)/P(X).

Posteriori=Likelihood\*Prior/Evidence.

Characteristics that led to consider this modelling technique are:

* They are easy to implement, give desirable results in most of the cases and performs well.
* If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data.
* A good bet if we want something fast and easy that performs pretty well.

Disadvantages: It cannot learn interactions between features. Also, it is stable with small data sets.

Dependent variable: **IsDenied**

Independent variables selected:AverageWeeklyWage + ClaimantAge\_at\_DOI + Gender + ClaimantType + IsHighRisk

**3.5 Linear regression**:

Linear regression is limited to predicting numeric output. In order for considering the next major cost driver, the total\_paid\_end, we decided to try this prediction method. Linear regressionimplements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, it shows optimal results.

The disadvantages of linear regression are being unable to handle outliers, multi-collinearity, and overfitting.

Dependent/target variable: **total\_paid\_end**

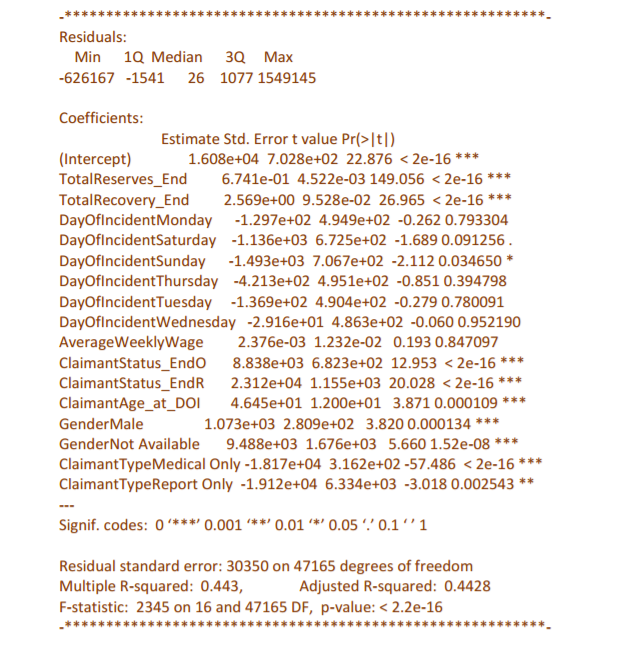
We are trying to estimate the total cost of the claim when the claim is generated. We initially considered a few independent input variables after looking at all columns. Then, a multiple linear regression model was built to get the estimated total paid.

The tool used was R and input variables are shown below:

lm(formula = TotalPaid\_End ~ TotalReserves\_End + TotalRecovery\_End + DayOfIncident + AverageWeeklyWage + ClaimantStatus\_End + ClaimantAge\_at\_DOI + +Gender + ClaimantType, data = dataset\_cl)

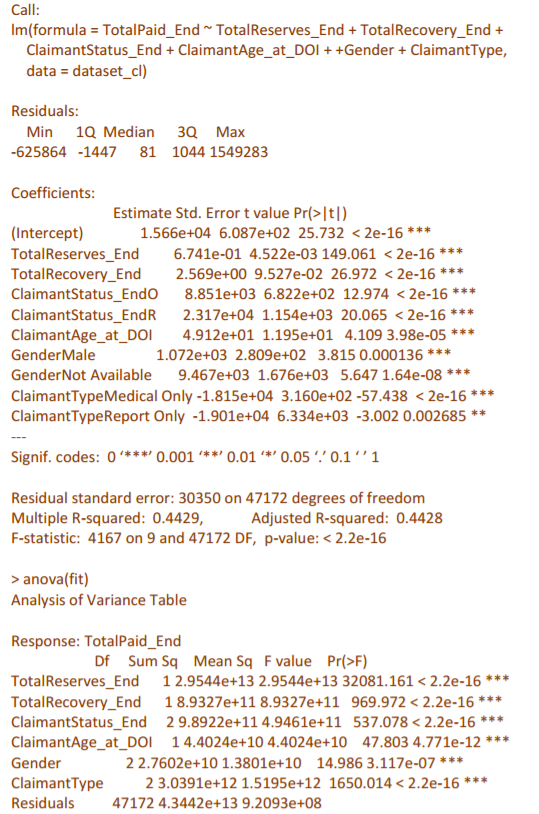
We got the below statistics (fig 3.2) for the above model

**Fig 3.2**



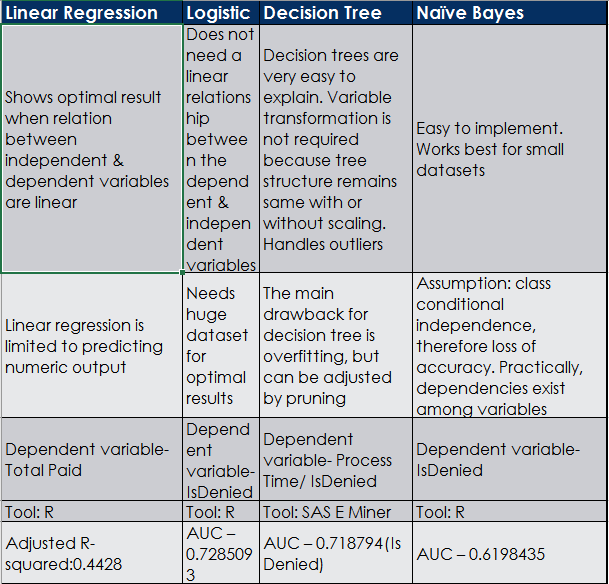
By looking at the results of the above model, we see that we have an R-square of 0.443 and Adjusted R-square of 0.4428. We looked at the p values of each variable and decided if all the variables were significant or if we could drop any of the variables. From observing the p-value of all the variables, we could see that the p value is more than 5% (0.05) for DayOfIncident and AverageWeeklyWage. Since, the model output shows that those 2 variables are not significant, we can drop those 2 variables and rerun our regression model. After our rerun, the results are shown in fig 3.3

**Fig 3.3**



From the above output, we see that the R-square and Adjusted R-square values are almost same even after dropping DayOfIncident and AverageWeeklyWage variables. Also, all the p-values seen are well below 5% (0.05), so all those variables are significant. Looking at the F-statistic, we can see that p-value is less than 5% which shows that model overall is a decent fit. One interesting observation to note is that the residual error is 30350. This is because there is a very high variation in the values of TotalPaid\_End.

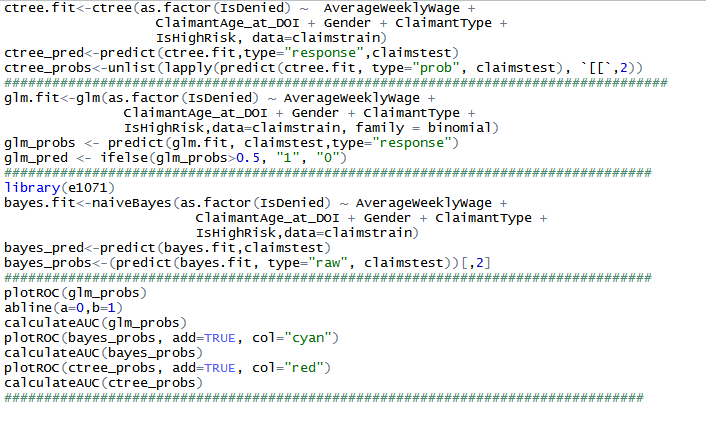
**3.6 Summary of Comparison:**

**Fig 3.4** 

**3.7 A Simple Model for Comparison of classifiers**

To compare the three classification methods, we implemented the same model with the classifiers- Naive Bayes, Decision tree and Logistic regression as shown in figure 3.5:

**Fig:3.5**

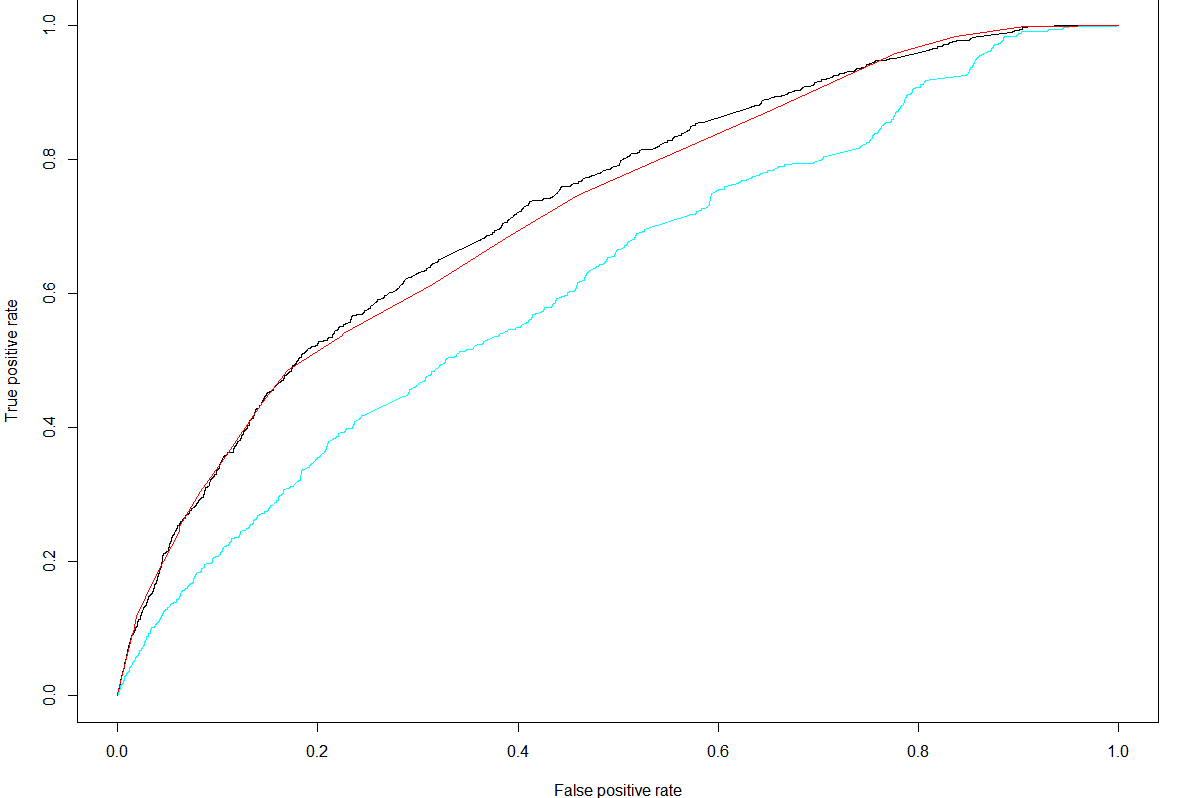


The ROC curve and Area Under the curve obtained are as shown in fig 3.6:

Black: Logistic regression (AUC:0.7285093)

Blue: Naive Bayes(AUC:0.6198435)

Red: Decision Tree(AUC:0.718974)

**Fig:3.6**

According to the ROC curve, logistic regression and decision tree are better techniques for our dataset. The team decided to proceed with decision tree, the details of our method is explained below.

**4. BUILDING THE PREDICTIVE MODEL:**

**Goal: Predict if a particular claim would take high, low or medium time to process**

STEP 1: CLASSIFICATION

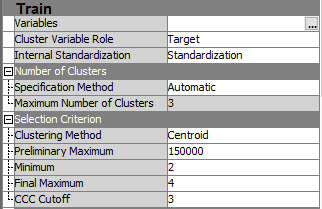
The *‘process\_time’* variable that we have is a numeric continuous variable. We needed to classify this variable into ‘high’, ‘low’, and ‘medium’ classes.

Trial 1: Clustering

We first tried clustering the ‘process\_time’ variable using the dataset given in SAS Eminer.

We asked for three clusters (since the range of the values was very large for just 2 - high and low) using the centroid/k-means clustering technique.

**fig 4.1**



The clusters we got were as follows:

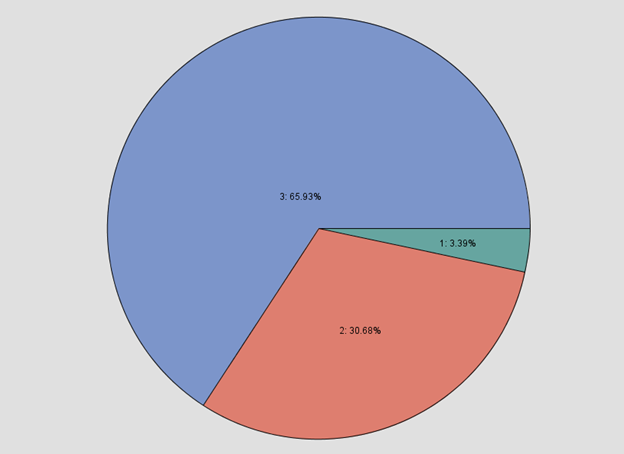
Cluster3: 1 to 1724 days (low) - 1 to 4.7 years

Cluster2: 1725 to 4857 days (med) – 4.7 to 13.3 years

Cluster1: 4865 to 9017 days (high) – 13.4 to 24.70 years

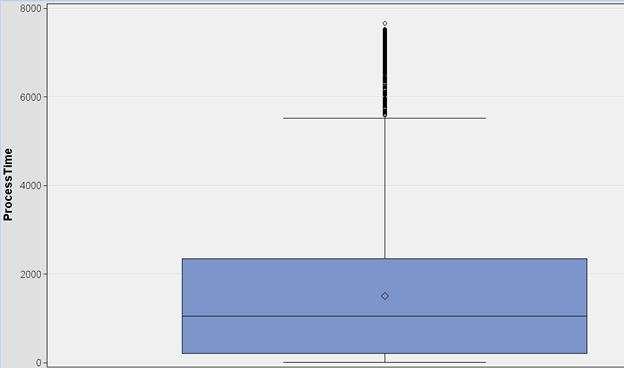
The distribution of the data in these clusters was as follows:

**Fig 4.2**



Now, it’s absurd that the lowest category has a range of 0 to 4.2 years, which is not low in itself. This clustering can be explained by the following distribution of the variable *‘ProcessTime’* in our dataset.

**Fig 4.3**



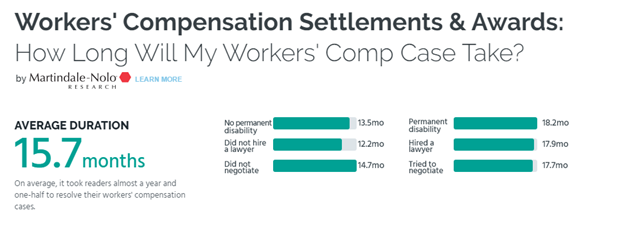
The clusters formed seem correct according to this, but do not fit our purpose since the low bucket itself has very high values. Thus, we must search for a new way to discretize the variable.

Trial2: External Data

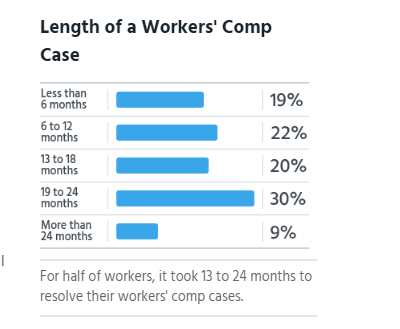
We researched online if there was any data available on how much time it takes for workers compensation claims to be settled. The website [www.lawyers.com](http://www.lawyers.com) has published the results of a research based on a survey of the time taken to settle workers compensation claims in the following link:

<http://workers-compensation.lawyers.com/workers-compensation-settlements-awards/how-long-will-workers-compensation-case-take-how-long-will-my-workers-compensation-case-take.html>

According to this resource, the average time it takes to settle a workers’ compensation claim is 15.7 months.



It also gives a breakdown in terms of percentages as follows:



Now, according to this data we clustered processing time variable as follows:

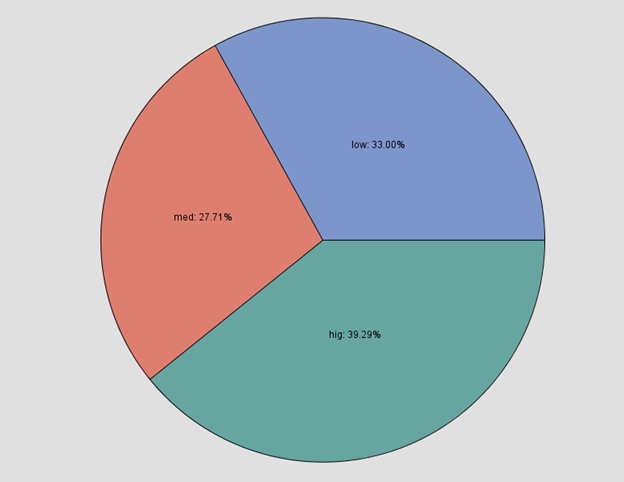
<1 year = low

1 year to 4 years= medium (extra grace period here of 2 years according to our data)

>4 years = high

According to this clustering, the distribution of the clusters is as follows:

**Fig 4.4**



STEP 2: IMPORT DATA

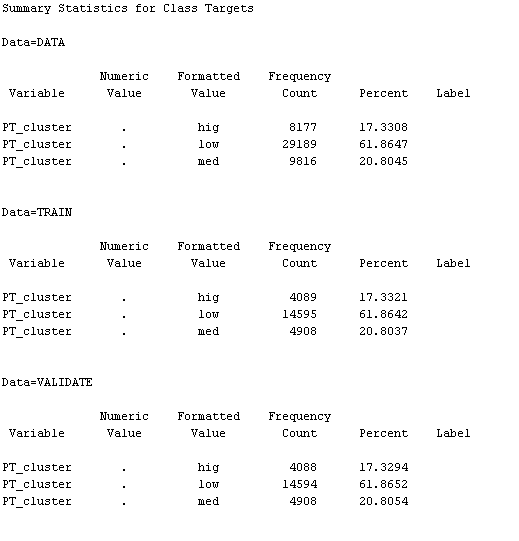
For the next step, we imported this data into SAS Enterprise Miner for modeling.

STEP 3: DATA PARTITION

To induce a decision tree and validate it, we needed to divide the data into test and validation datasets.

The data was randomly divided into two equal halves for this purpose as can be seen from the below screenshot.

**Fig 4.5**



STEP 3: DECISION TREE INDUCTION

Next, for the induction of decision trees, we took the following variables as parameters:

1. AverageWeeklyWage
2. Body\_Part
3. BodyPartRegion
4. ClaimantAge\_at\_DOI
5. ClaimantStatus\_End
6. ClaimantType
7. ClaimID
8. DayOfIncident (Day of the Week)
9. Gender
10. Injury\_Nature

The Target variable is PT\_cluster which is nothing but the ‘ProcessTime’ variable classified as ‘high’, ‘medium’ or ‘low’.

We set a few conditions as follows:

Maximum branches = 2

Maximum depth = 9 (to allow more complexity)

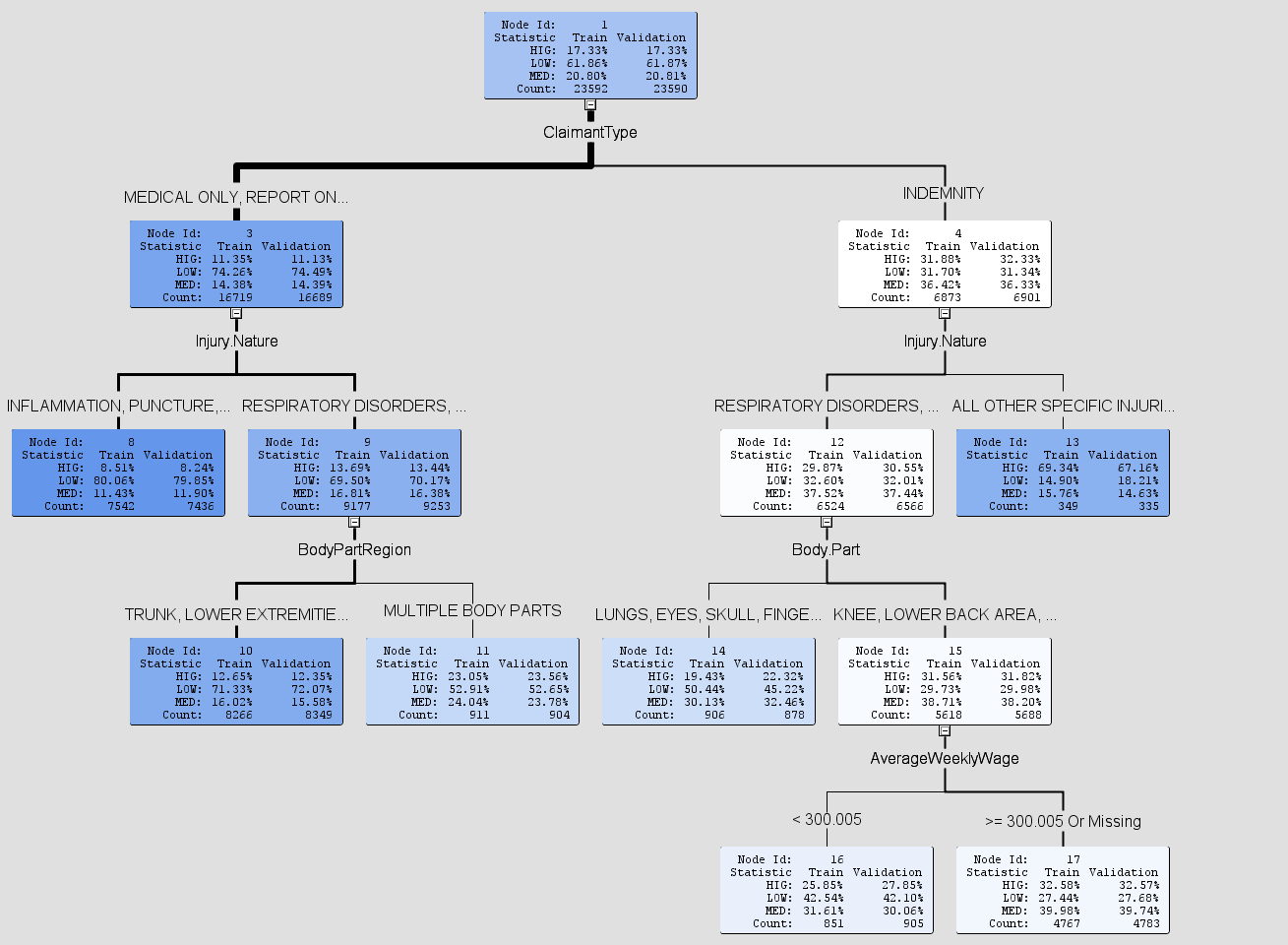
Leaf size = 240 (around 0.5-1% of the dataset size is standard)

Assessment measure = decision

Next, we induced the decision tree interactively so that we have the control to stop in case there isn’t any further improvement happening. This helps us reduce complexity and gives some control. 20000 records are randomly used for this.

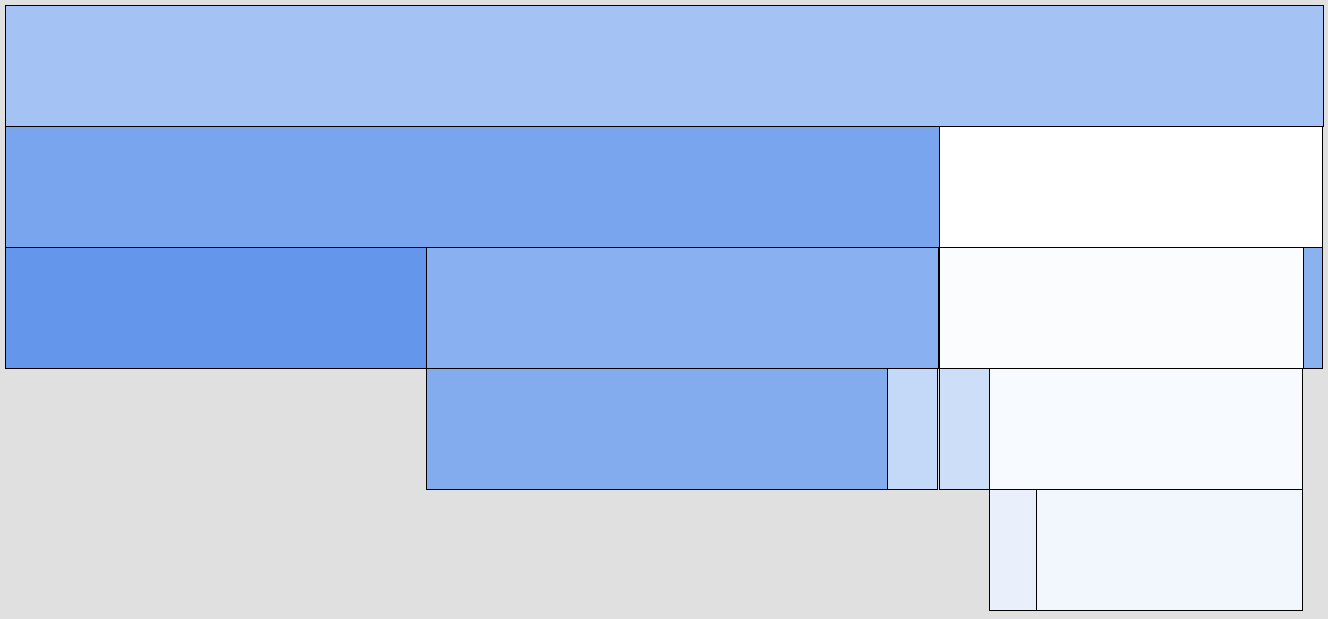
Once we have the induced structure, the ‘Decision Tree’ node is then run for the entire data. The final tree is attached below.

**Fig 4.6**



The tree structure can also be interpreted from the below tree map:

**Fig 4.7**



The darker the color, the better the data is divided into the categories ‘high’, ’medium’ and ‘low’.

As can be seen from the lowest layer, most of the data is well divided.

The misclassification rate and the average squared error for both the training dataset and the validation dataset is quite small for this model.

**Fig 4.8**





Therefore, if the claims company made use of this model, they can identify which of their future claims will tend to take a higher time based on the past data. This will help them concentrate resources in such cases, thus reducing processing time. Overall, it could help them improve their position in the market as well.

**5. INSIGHTS**

5.1 Insights from the visuals:

1. According to fig 2.1, more incidents are reported on Wednesdays(middle of the week) and the least during the weekends. This is different from what we originally hypothesized; we thought that the highest amount of incidents would be reported on Monday due to the people doing unworkrelated things during the weekend, get injured, then file for a claim on Monday.

2. Fig. 2.3 shows that the maximum offdays (difference between claims closed date and claims open date) is seen for the workers reported with mental disorders and other psychological problems .

3.The trend of cost drivers (Claims process time and Total amount paid) of Workers Compensation Claims Processing company was explored. From 2001, the average total paid and average process time are consistently remaining low. However, the average process time tends to increase from 2011(From fig 2.4)

5.2 Insights from Predictive Modelling (fig: 4.6)

* The major attribute that caused the split in our decision tree is Claimant type. It shows that 32% of indemnity claims are at high risk. Also 11% of medical/report only cases are at high risk.
* Injury nature is a very important attribute in deciding the claims processing time.
* All indemnity claims having reported with “all specific injuries” will have higher possibility to become high risk claims
* Indemnity claims with upper extremity illnesses are prone to be high risk.
* The claims reported with respiratory problems must be seen whether it is indemnity or medical only. If it is medical, then the severity of cost due to process time is lesser compared to that of the medical type.

* Any indemnity claims having higher average weekly wage (>=300) has higher possibility to take more process time

**6. RECOMMENDATIONS:**

* The company needs to see the chances of fraud. They need to determine whether the employees are getting injured during the weekends and they are waiting until the middle of the week to report it for getting the workers compensations (Insight 1) or if they are truly getting hurt at work.
* The company can offer periodic counselling for the employees, so that the work related mental stress can be addressed. Diagnosis of mental illness is subjective. Hence the company can ask the doctors for specific reasons (root causes) for his/her diagnosis (Insight 2).
* Since the process time tends to increase from 2011, more analysis on the process time is required to minimize the cost (Insight 3). We used this information to make our predictive model.
* Insights from the predictive modelling suggest that based on the model, the future claims can be classified into high, low or medium. From the insights gathered, we would recommend to assign cases depending on the results from the predictive modelling. For example, those with high process time can be assigned with experienced adjuster and additional resources.
* Work related disorders like Carpel tunnel syndrome (fig 2.2) can be avoided to some extent by providing better precautionary measures like ergonomic keyboard or mouse. By providing employees with ergonomic setup such cases will highly decrease the risk and cost for insurance company.

**7. ANALYICS 3.0**

1. Building Analytics team & create “chief analytics officer” role: Analytics team should be formed in Workers compensation Claims Company. Assembling competitive Data Scientists & Analysts and developing the skilled employees is required for Analytics work. The company needs to create “chief analytics officer” role to superintend the building and use of analytical capabilities.  
  
2. Creating analytics products: The Company’s management must try to compete on analytics not only in the traditional sense (by improving internal business decisions) but also by creating more-valuable products and services.   
  
3. Implementing “Agile” analytical methods: New “agile” analytical methods and machine-learning techniques should be implemented to produce insights at a much faster rate. Like agile systems development, these methods involve frequent delivery of partial outputs to the project stakeholders. The challenge in the 3.0 era is to adapt operational, product development, and decision processes to take advantage of what the new technologies and methods can bring forth.