ITCS-6100

Big Data Analytics for Competitive Advantage

Workers Compensation Claims Project Overview

Presented by

- Diptesh Nath
- Jagadeesh Ballur
- Manasa Puli
- Pooja Sharma
- Yagnesh Vadalia

Contents

Executive Summary

The problem background and data exploration of the claims data were completed earlier and that this report is focused on the modeling and recommendations.

key findings from the data visualization of the merged dataset:

1. Upper Extremities has the highest number of claims. The Bill review costs is the highest for the trunk body part region.
2. Considering the weekday of the incident, freezing injury has incidents on Tuesday, Wednesday and maximum on Saturday. Another is Burns which was highest number of claims on Thursdays.
3. low risk claims are related with relatively lower number of days to closure and high risk claims are associated with high number of days to closure.
4. Strain and Contusion injuries are dominating the claim cost drastically.

key findings from predictive modeling:

Linear Regression:

Linear regression needs a continuous target variable and continuous predictor variables. Since we are aiming to reduce costs and if we can predict the transaction costs of the claims which are used for classifying high and low risk claims then it will serve our purpose.

Clustering:

k -means clustering on total paid variable to classify the claims into low and high risk claims. After the initial cluster analysis, we further reclassified certain claims as high risk claims because although their individual claim costs were low however those claims collectively were having the highest claim costs. Example: claims with injury nature as Strain, Sprain, Concussion and Contusion had very high costs.

Random tree model

Random tree on the claims data to see how the tree splits the data into high and low risk claims. The dependent variable was risk class with high and low values. One of the highlights from the Random tree model is the combination of wage and age determined the risk category in a number of places

Logistic Regression

The claims dataset has a good mix of continuous and categorical variables and a logistic regression is a good choice of model to predict the claim risk as high or low using both continuous and categorical variables. Weka Tool is used to run Logistic Regression.

Recommendations:

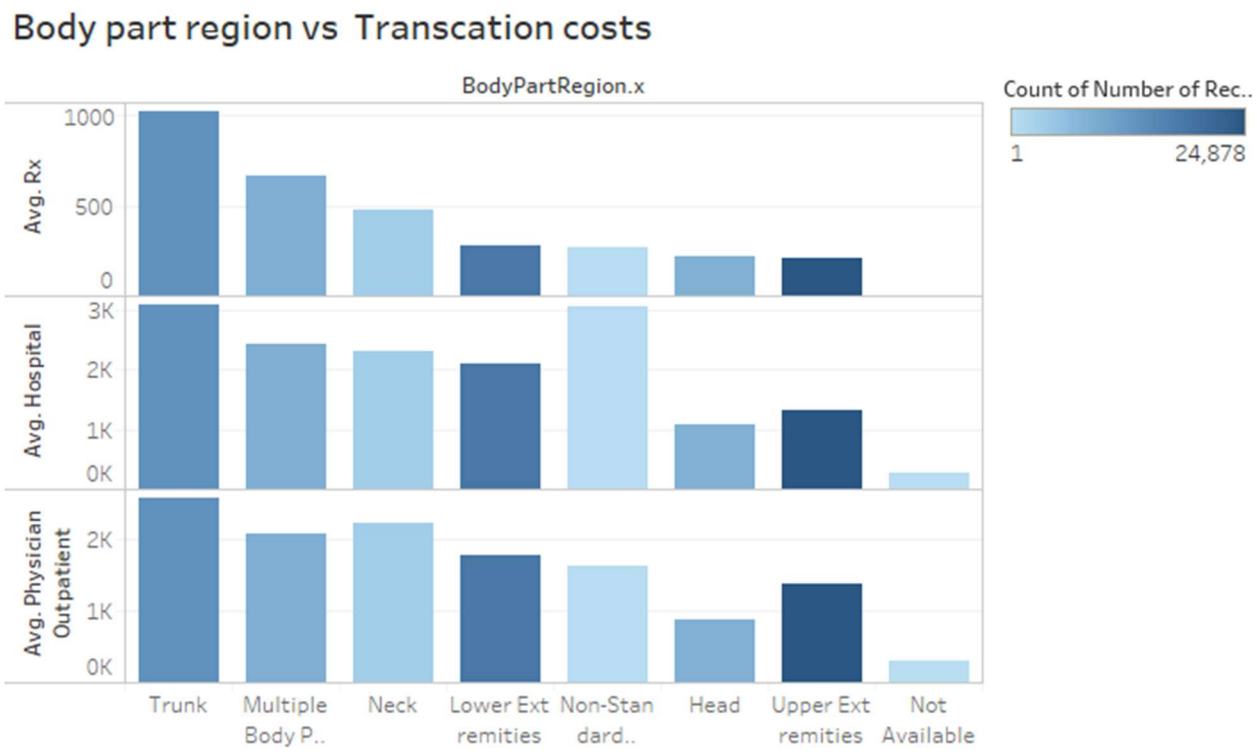
1. The company can come up with plans focused on weekdays and injury nature to reduce costs.

2. The company should not ignore the claims which have low costs as low risks because eventually it won't bring down the overall claim costs if such claims are ignored.

3. From the odds ratio in logistic regression, it is evident that claims with injury nature as Sprain are 7 times more likely to be high risk claims compared to other claims. This will help the company to pay more attention to such claims and take measures to mitigate the occurrence of such injuries and in turn reduce the company's overall claim costs.

Data Visualizations

Visualization 1: Bill Review Costs Vs Body Part Region



Viz 1

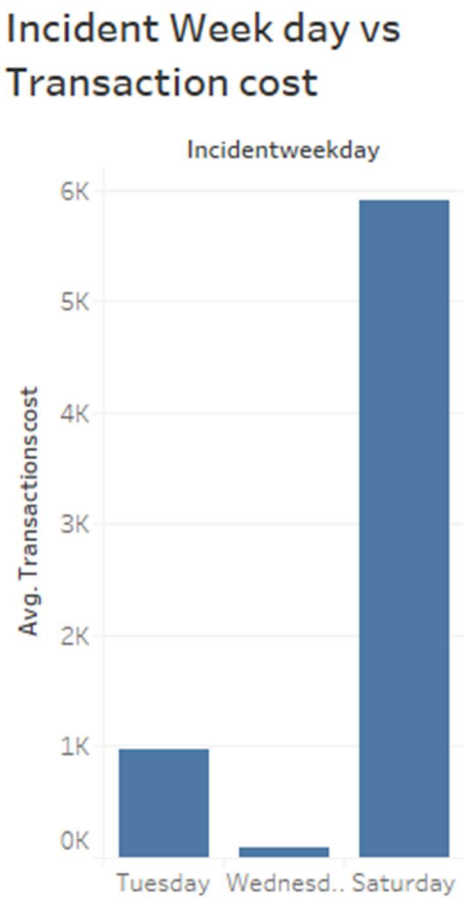
In the above visualization each bar represents one body part region. Upper Extremities have more number of claims , indicated by darkest color. The Bill review costs is the highest for the trunk body part region.

-

Visualization 2: WeekDay Vs Injury Nature

The below visualization has a filter on injury nature and has only “Freezing”injury nature on the filter. There are some injuries which occur on specific weekdays. For example - freezing injury has incidents on Tuesday, Wednesday and maximum on Saturday. Another example not shown in the below chart is Burns - which was highest on Thursdays.

The company can come up with plans focused on weekdays and injury nature to reduce costs.



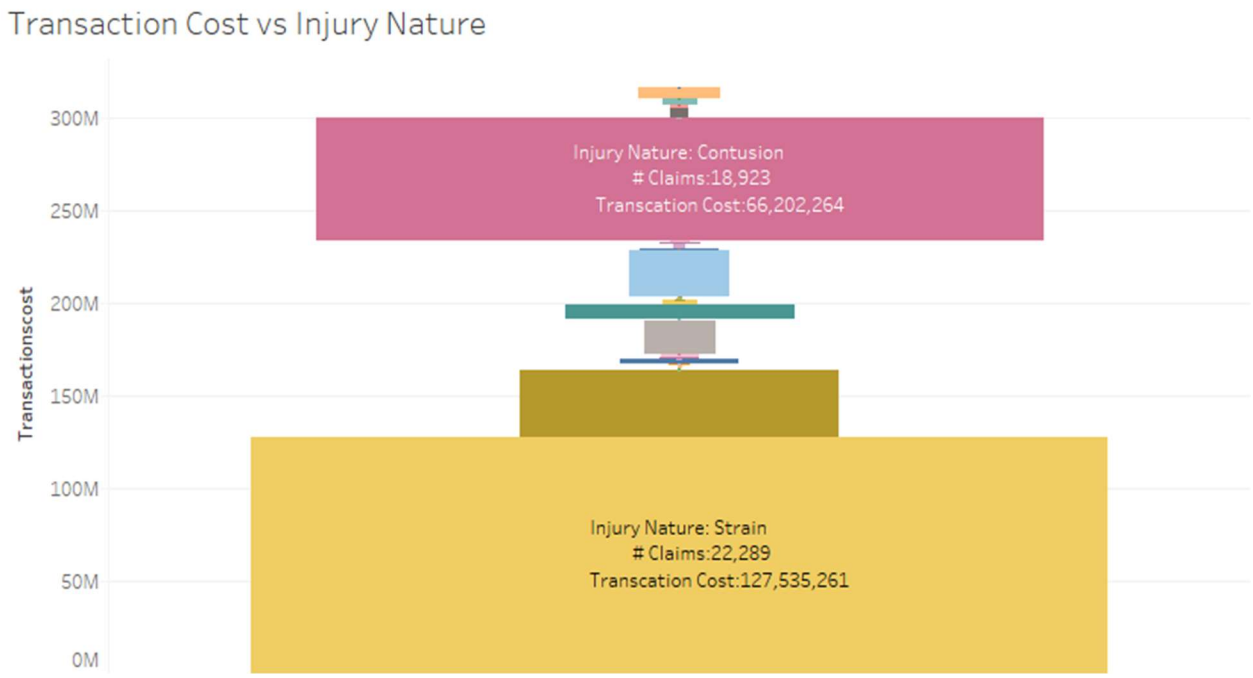
Visualization 2

-

Visualization 3:

In this visualization we have explored the various transactions costs - hospital, physical outpatient and Rx and compared that with injury nature. We have reduced the 50 plus injuries into 6 injury categories to get a more precise chart. The subsequent sections have more details.

.



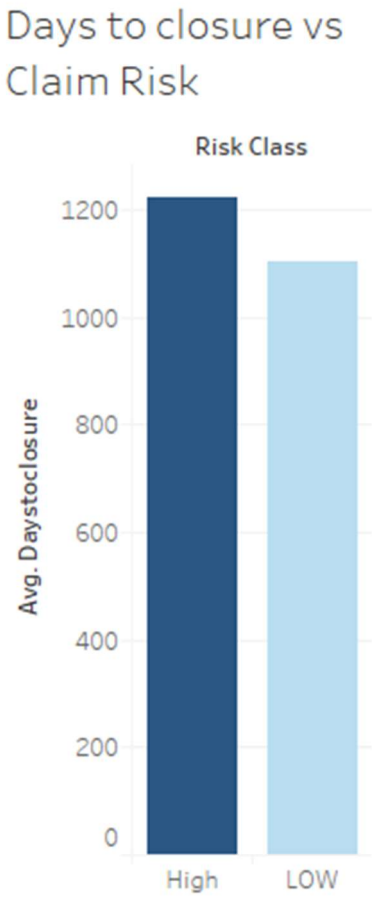
Visualization 3

Strain and Contusion injuries are dominating the claim cost drastically.

-

Visualization 4:

In the chart we have used that classification variable and explored the number of days to closure. The high risk claims have a darker color and taller bar compared to low risk claims indicating that high risk claims are associated with high number of days to closure.



Visualization 4

The high risk claims have a darker color and taller bar compared to low risk claims indicating that high risk claims are associated with high number of days to closure

---

-

New Variables

Risk Rating/Risk Class: The claims are classified into two types based on the costs. The Low risk claims are taken as 0 and the High risk claims are taken as 1. We implemented k-means

clustering to classify the claims into high risk or low risk. Initially The clustering is done based on the aggregate of the costs for each injury nature. The riskrating variable uses 0 and 1 notation and RiskClass variable uses High and Low notation to represent claim risks.

Injury Nature Category: The injury nature has many classifications. We reduced the injury nature into less number of categories. From the analysis we did in part 1 - apart from Strain, Sprain, Fracture, Contusion, Multiple Physical Injuries and Laceration , the other injury natures' were having similar impact on the total paid costs of the claims so we clubbed these into a new category called "Others". So, there is an InjuryNatureCategory column which has reduced injury natures - namely - Strain, Sprain, Fracture, Contusion, Multiple Physical Injuries, Laceration and Others.

Incidentweekday: Extracted the day of the week on which the incident occurred using the incident date given in the original dataset. We used date packages in R language to extract this weekday. This was done in account to see whether there are any injuries that occur throughout the week or concentrated particular to any day, so that it may relate to any particular job those employees are concerned with. In addition, this variable also support the claim made in previous assumption of splitting into injury categories.

DaysToClosure: For closed claims, we subtracted the number of days from the claim open date and claim closed date to get the number of days it took to close each claim.

UpperWageLimit: We binned the continuous average weekly wage variable into bins such as 0 to 200 weekly wage as 200, 200 to 400 weekly wage as 400 and so on. The reason is we want to try decision trees and a binned variable will be more effective compared to a continuous range for that model.

UpperAgeLimit: Similar to the upperwage limit variable, we also binned the age variable into groups. such that 15 to 20 and 20 to 30 and 30 to 40 and so on. The age group is more effective than individual age values for a classification model to use as an independent variable.

Comparative Analysis of Predictive Modelling Techniques

Clustering:

The objective of our project is to reduce costs for the claims company. In order to identify the high cost claims from the low cost claims we introduced a new variable called claim risk. We used k -means clustering on total paid variable to classify the claims into low and high risk claims. We use clustering since, the algorithm itself determines the two centroids corresponding to the dataset, such that making the classification more centric to the dataset presented, rather than going for the market trends.

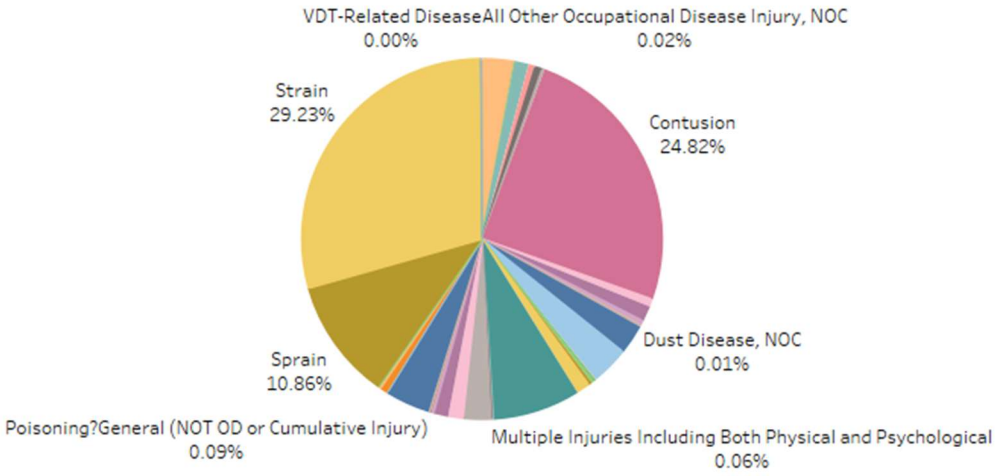
Below is the code that we used to cluster the claims and then assign a claim category to each of the claims in order to use it in subsequent models for predictions.

```
library(cluster)
claimclusters <- kmeans(latest$TotalPaid_End, centers = 2)
whichbin <- claimclusters$cluster;

table(whichbin)

bin1<- rownames(latest[ claimclusters$cluster==1,])
bin2 <- rownames(latest[ claimclusters$cluster==2,])

bin1meanpaid <- mean(latest$TotalPaid_End[which(rownames(latest) %in% bin1)])
bin2meanpaid <- mean(latest$TotalPaid_End[which(rownames(latest) %in% bin2)])

tempCat <- c("-1", "-1")
maxMean <- max(bin1meanpaid,bin2meanpaid)
minMean <- min(bin1meanpaid,bin2meanpaid)

if(bin1meanpaid == maxMean){
  tempCat[1]="HIGH"
}else if(bin1meanpaid == minMean){
  tempCat[1]="LOW"}

if(bin2meanpaid == maxMean){
  tempCat[2]="HIGH"
}else if(bin1meanpaid == minMean){
  tempCat[2]="LOW"}


for(i in 1:nrow(latest)){
  if(whichbin[i]==1){
    latest$Risk[i] <- tempCat[1]
  }
  else if(whichbin[i]==2){
    latest$Risk[i] <- tempCat[2]
  }
}

low_claims<- latest[which(latest$Risk%in%"LOW"),]
high_claims<- latest[which(latest$Risk%in%"HIGH"),]

write.csv(latest, file = "latest_classified.csv",row.names=FALSE)
```

After the initial cluster analysis, we further reclassified certain claims as high risk claims because although their individual claim costs were low however those claims collectively were having the highest claim costs. For example claims with injury nature as Strain, Sprain, Concussion and Contusion had very high costs. So the company should not ignore these claims as low risks because eventually it won't bring down the overall claim costs if such claims are ignored.

This is shown in the image below:

support_1



Here we can see that sprain and contusion alone accumulates nearly 54%.

The primary objective of this project is to predict the claims that might cost them heavily, which are classified as *high risk claims or equivalent to 1 in terms of a binary variable.* For this we considered 8 predictor variables which we believed was having higher impact with the target variable, from the exploratory data analysis. These are :

- Gender
- Claimant type
- Body part region
- Injury nature Category
- Transaction cost
- Upper age limit

- Upper wage limit

With the target variable being *Risk Class* or *Risk rating* depending on the model.

Linear Regression

Linear regression needs a continuous target variable and continuous predictor variables. Since we are aiming to reduce costs and if we can predict the transaction costs of the claims which are used for classifying high and low risk claims then it will serve our purpose.

We tried a simple linear regression to predict the transaction costs of the claims which contribute to the total costs using days to closure(new variable) and claimant age (without binning the age) as the independent variables.

Below is the code we used

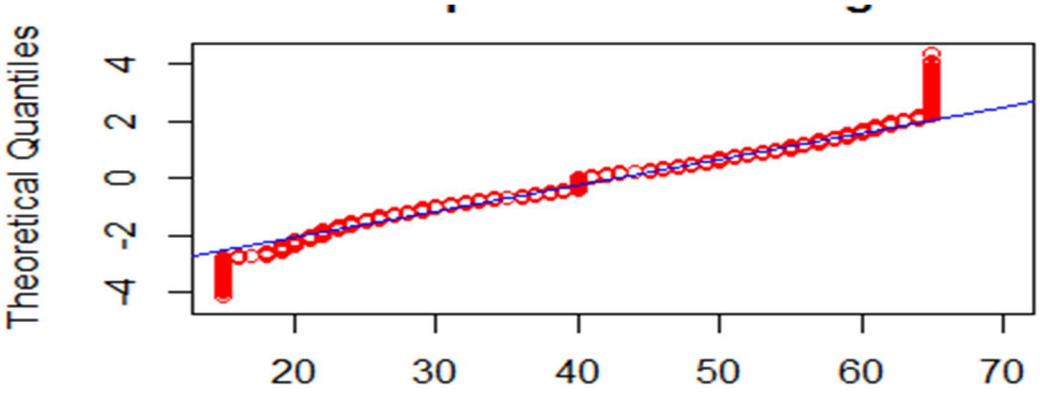
```
lm.out <- lm(latest$Transactionscost ~ latest$daystoclosure + latest$ClaimantAge_at_DOI)
summary(lm.out)
plot(lm.out)
```

Here are the linear regression results:

```
Call:
lm(formula = latest$Transactionscost ~ latest$daystoclosure +
    latest$ClaimantAge_at_DOI)

Residuals:
    Min      1Q  Median      3Q     Max
 -35269   -3330    -970     357 1113886

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.488e+03  2.793e+02  -8.907   <2e-16 ***
latest$daystoclosure     3.363e+00  3.221e-02 104.416   <2e-16 ***
latest$ClaimantAge_at_DOI  6.357e+01  6.430e+00   9.887   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19200 on 76248 degrees of freedom
Multiple R-squared:  0.1281,    Adjusted R-squared:  0.1281
F-statistic:  5601 on 2 and 76248 DF,  p-value: < 2.2e-16
```

We got a 12% adjusted R square value which was not very great. Also when we did the residual plots for this model, it did not follow the linear regression assumptions. The plot for age was normal except for at the boundary values. This is explainable given the fact that we cleaned the age column and replace all ages less than 15 with 15 and all ages greater that 65 with 65.

Below is the qq plot for age:



Visualization 5: *QQ- plot for Claimant Original age at the time of filing*

An adjusted R-squared of 12% is not good enough and we know that other factors such as injury nature do play an important part in claim costs so we tried predictive modelling using decision trees and logistic regression in the subsequent sections.

Random Forests

We tried random forests to predict the claim risk factor high or low using the binned age, wage, days to closure and other factors. We created binned variables especially for classification models so that we get better classification results.

We ran a random forest algorithm in Weka Tool to get predictions on claim risk category i.e. whether the claims are high risk or low risk claims.

We used seven attributes for random forest unlike the linear regression where we were able to use only a couple of attributes. The seven attributes we used for random forests are:

Gender - 0 or 1 where 0 stands for male and 1 stands for female. We did this encoding so that it will help in logistic regression also, Injury Nature, BodyPart Region, Claimant Type - Indemnity or Medical, Upper wage limit - explained in “New Variables” section, Upper Age Limit - explained in “New Variables: section and Risk Class - as the target variable.

The dataset was divided into 66% training dataset and remaining as test dataset.

We got high accuracy results using this technique.However we are skeptical regarding the results because it could be due to imbalanced dataset since there are more high risk claims compared to low risk claims.

The results from the random forest we ran in Weka tool are as follows:

```
Instances:    76251
Attributes:   9
              Gender
              ClaimantType
              InjuryNature_category
              BodyPartRegion.x
              daystoclosure
              Transactionscost
              Upperwagelimit
              Upperagelimit
              RiskClass
Test mode:    split 70.0% train, remainder test

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner
```

Below is the evaluation on the test data which was 30% of the total dataset.

```
=== Evaluation on test split ===

Time taken to test model on training split: 0.6 seconds

=== Summary ===

Correctly Classified Instances       22851               99.8951 %
Incorrectly Classified Instances        24                0.1049 %
Kappa statistic                          0.9978
Mean absolute error                      0.0018
Root mean squared error                  0.0283
Relative absolute error                  0.3779 %
Root relative squared error              5.8503 %
Total Number of Instances            22875

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.001    0.999      0.999    0.999      0.998    1.000     1.000     High
                 0.999    0.001    0.998      0.999    0.999      0.998    1.000     1.000     LOW
Weighted Avg.    0.999    0.001    0.999      0.999    0.999      0.998    1.000     1.000

=== Confusion Matrix ===

     a     b   <-- classified as
 14280    13 |     a = High
    11  8571 |     b = LOW
```

From the random forests we got more accurate results compared to the linear model. As is evident from the confusion matrix 14280 claims were correctly classified as high risk and 8571 as low risk claims. We achieved an accuracy of above 90%.

Random Trees

We also ran a random tree on the claims data to see how the tree splits the data into high and low risk claims. The dependent variable was risk class with high and low values. The predictor variables were the below attributes -

Gender
Claimant type
InjuryNature_Category
Body Part Region
daystoclosure
Upper Wage Limit
Upper Age Limit

Below is the actual tree diagram for the initial few splits-

```
BodyPartRegion.x = Trunk
|   Upperagelimit < 45
|   |   InjuryNature_category = Strain : High (4631/0)
|   |   InjuryNature_category = Fracture : LOW (55/0)
|   |   InjuryNature_category = Contusion : High (1001/0)
|   |   InjuryNature_category = Other
|   |   |   Gender < 0.5
|   |   |   |   Upperwagelimit < 500 : LOW (60/0)
|   |   |   |   Upperwagelimit >= 500
|   |   |   |   |   daystoclosure < 2347 : LOW (268/0)
|   |   |   |   |   daystoclosure >= 2347
|   |   |   |   |   |   daystoclosure < 2478.5 : High (1/0)
|   |   |   |   |   |   daystoclosure >= 2478.5 : LOW (11/0)
|   |   |   Gender >= 0.5
|   |   |   |   daystoclosure < 2424
|   |   |   |   |   ClaimantType < 1.5 : LOW (211/0)
|   |   |   |   |   ClaimantType >= 1.5
|   |   |   |   |   |   Upperagelimit < 35
|   |   |   |   |   |   |   daystoclosure < 718 : LOW (6/0)
|   |   |   |   |   |   |   daystoclosure >= 718
|   |   |   |   |   |   |   |   daystoclosure < 1028.5 : High (1/0)
|   |   |   |   |   |   |   |   daystoclosure >= 1028.5 : LOW (2/0)
|   |   |   |   |   |   Upperagelimit >= 35 : LOW (26/0)
|   |   |   |   daystoclosure >= 2424
|   |   |   |   |   daystoclosure < 2598 : High (1/0)
```

The full size of the tree: 1020

Below are the results of the model on the test data:

```
=== Summary ===

Correctly Classified Instances       22814               99.7333 %
Incorrectly Classified Instances        61                0.2667 %
Kappa statistic                          0.9943
Mean absolute error                      0.0026
Root mean squared error                  0.0475
Relative absolute error                  0.5553 %
Root relative squared error              9.8049 %
Total Number of Instances            22875

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.004    0.997      0.998    0.998      0.994    0.998     0.998     High
                 0.996    0.002    0.997      0.996    0.996      0.994    0.998     0.996     LOW
Weighted Avg.    0.997    0.003    0.997      0.997    0.997      0.994    0.998     0.997

=== Confusion Matrix ===

     a     b   <-- classified as
 14268    25 |     a = High
    36  8546 |     b = LOW
```

Some of the highlights from the Random tree model were:

A combination of wage and age determined the risk category in a number of places. One of the split is shown below -

```
|   |   ClaimantType >= 1.5
|   |   |   Upperwagelimit < 700
|   |   |   |   Gender < 0.5
|   |   |   |   |   Upperagelimit < 45
|   |   |   |   |   |   Upperagelimit < 25 : LOW (1/0)
|   |   |   |   |   |   Upperagelimit >= 25
|   |   |   |   |   |   |   Upperwagelimit < 300
|   |   |   |   |   |   |   |   InjuryNature_category = Strain : High (0/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Fracture : High (0/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Contusion : High (13/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Other : LOW (12/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Multiple Physical Injuries Only : LOW (2/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Sprain : High (0/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Laceration : High (4/0)
|   |   |   |   |   |   |   |   InjuryNature_category = All Other Specific Injuries, Noc : High (0/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Concussion : High (0/0)
|   |   |   |   |   |   |   Upperwagelimit >= 300
|   |   |   |   |   |   |   |   InjuryNature_category = Strain : High (0/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Fracture : LOW (6/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Contusion : High (38/0)
|   |   |   |   |   |   |   |   InjuryNature_category = Other
|   |   |   |   |   |   |   |   |   daystoclosure < 6174 : LOW (10/0)
|   |   |   |   |   |   |   |   |   daystoclosure >= 6174 : High (3/1)
|   |   |   |   |   |   |   |   InjuryNature_category = Multiple Physical Injuries Only : LOW (1/0)
```

Logistic Regression

The claims dataset has a good mix of continuous and categorical variables and a logistic regression would be a good choice of model to predict the claim risk as high or low using both continuous and categorical variables. A detailed analysis on how we used the logistic regression is present in the following section - “Predictive Modelling Analysis”.

Compared to Random forests, for logistic regression we took a lesser proportion of high risk claims to ensure that we are not facing the problem of imbalanced dataset. In order to achieve that we reassigned the risk category of certain claims from high to low specifically the claims where injury nature was Strain and Contusion.

We also tried Logistic Regression to make our final predictions on the risk of claims and predicting whether the claims will be high risk or low risk in terms of costs involved.

We fed our merged claims dataset into the Weka Tool and ran Logistic Regression on it. Our target variable was the risk class and following were the independent variables we fed to the model -

```
Attributes:   8
              Gender
              ClaimantType
              InjuryNature
              BodyPartRegion
              daystoclosure
              Upperwagelimit
              Upperagelimit
              RiskClass
Test mode:    split 70.0% train, remainder test
```

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...

Variable	Class High
Gender	0.203
ClaimantType	109.4586
InjuryNature=Strain	114.2164
InjuryNature=Fracture	-365.5047
InjuryNature=Contusion	69.4402
InjuryNature=Other	-236.8268
InjuryNature=Multiple Physical Injuries Only	-394.1324
InjuryNature=Sprain	107.9561
InjuryNature=Laceration	237.7374
InjuryNature=All Other Specific Injuries, Noc	-424.0727
InjuryNature=Concussion	562.6852
BodyPartRegion=Trunk	0.5298
BodyPartRegion=Upper Extremities	-0.4327
BodyPartRegion=Lower Extremities	-0.2942
BodyPartRegion=Multiple Body Parts	1.0319
BodyPartRegion=Head	-0.4102
BodyPartRegion=Neck	0.2974
BodyPartRegion=Non-Standard Code	-0.538
BodyPartRegion=Not Available	49.893
daystoclosure	0.0004
Upperwagelimit	0.0005
Upperagelimit	-0.0015

Below were the odds ratio from the logistic regression results -

Odds Ratios...

Variable	Class High
Gender	1.2251
ClaimantType	3.445688149173233E47
InjuryNature=Strain	4.013806478777405E49
InjuryNature=Fracture	0
InjuryNature=Contusion	1.437167728810751E30
InjuryNature=Other	0
InjuryNature=Multiple Physical Injuries Only	0
InjuryNature=Sprain	7.669185781671101E46
InjuryNature=Laceration	1.7702249311308348E103
InjuryNature=All Other Specific Injuries, Noc	0
InjuryNature=Concussion	2.3501578122124278E244
BodyPartRegion=Trunk	1.6986
BodyPartRegion=Upper Extremities	0.6488
BodyPartRegion=Lower Extremities	0.7452
BodyPartRegion=Multiple Body Parts	2.8065
BodyPartRegion=Head	0.6635
BodyPartRegion=Neck	1.3463
BodyPartRegion=Non-Standard Code	0.5839
BodyPartRegion=Not Available	4.658596883926986E21
daystoclosure	1.0004
Upperwagelimit	1.0005
Upperagelimit	0.9985

As we can see from the odds ratio, it is evident that claims with injury nature as Sprain are 7 times more likely to be high risk claims compared to other claims. This is true because is one of the top injury nature that is driving the claims costs and classifying such claims as high risks will help the company to pay more attention to such claims and take measures to mitigate the occurrence of such injuries and in turn reduce the company's overall claim costs.

All the above models we used earlier, injury nature category was used in classifying, resulted in a high accurate predictions. This support our view on the other predictor variables used, that they are significant in classifying the risk. So for our analysis we use C5.0 model without injury category variable in the predictors and try to get a model that best serves the purpose.

Predictive Modelling Analysis

As said earlier, we have used C5.0 model, as our main model in predicting the risk claims and thereby come for a decision to be taken in order to reduce cost involved in compensation

A brief note on C5.0 model : C4.5 algorithm employs divide-and-conquer strategy in constructing the decision tree. For a given set S of case it develops the algorithms as follows:

If all the cases in S belongs to same class or S is small, the tree is leaf labelled with most frequent class in S. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2. According to the outcome for each case, and apply the same procedure recursively to each subset2.

We used the below variables as dependent variable: Risk Class - high or low.

We used the below variables as predictor variables:

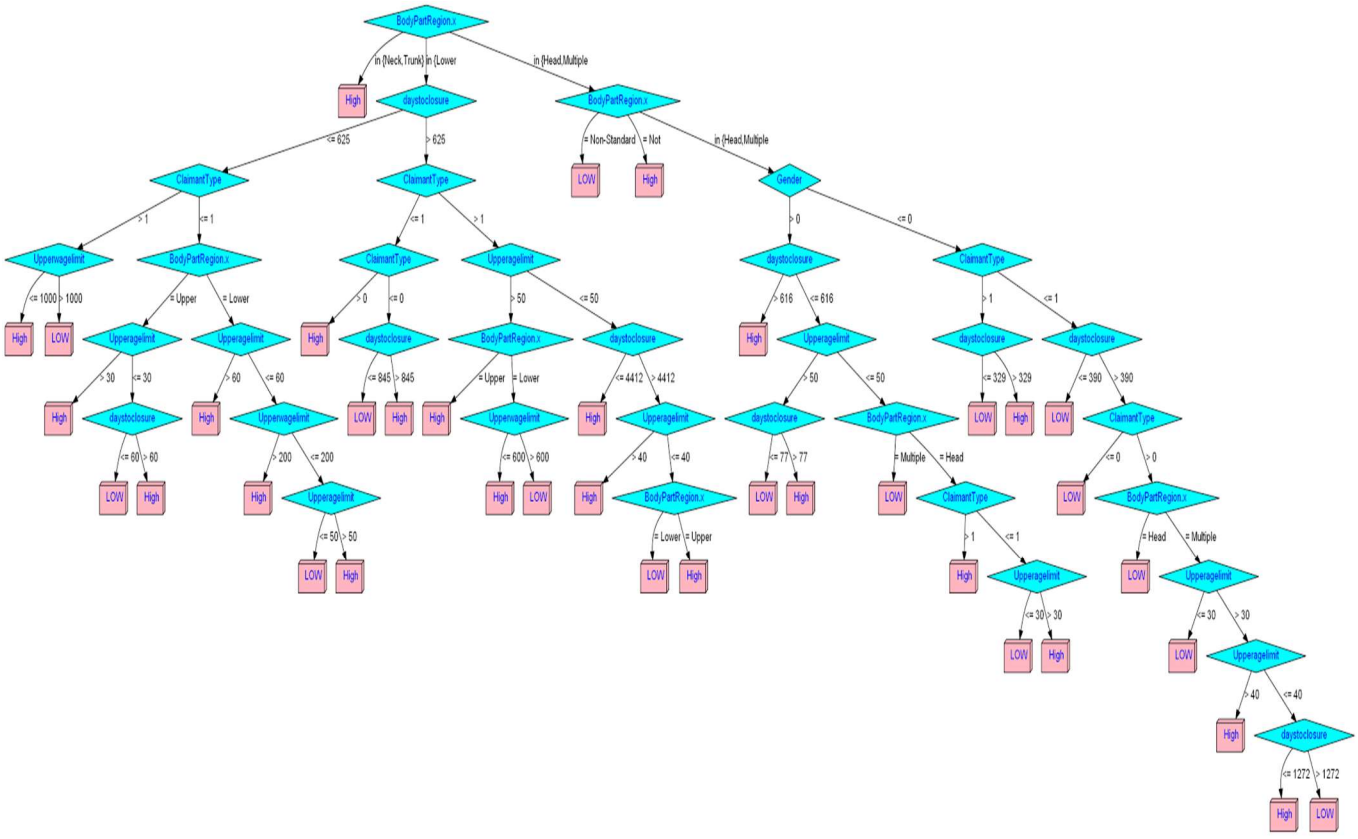
1. Gender
2. Claimant Type
3. Body Part Region.x
4. daystoclosure
5. Upper Wage Limit
6. Upper Age Limit

We ran the C5.0 in R studio and below are the details :

```
# Using C5.0 for modelling

#Training the model

x<-data_train[,c(5,6,8,9,15,16)]

y<-data_train$RiskClass

y<-as.factor(y)

library(C50)

data_model_c50 <- C5.0(x,y)

data_model_c50

summary(data_model_c50)

# Testing the model using Test data set

data_predict_c50 <- predict(data_model_c50,data_test)

# Validating the accuracy of the model using chi-square test for proportions
```

```
library(gmodels)
CrossTable(data_test$RiskClass, data_predict_c50,
        prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
        dnn = c('actual ', 'predicted '))
```

Below is the plot of the decision tree built by the C5.0 algorithm we used above.



* *complete image in the submission document*

We have also included the above tree plot as a separate png image because it will be more zoomed out when viewed as a separate image.

Total Observations in Table: 22875

actual	predicted High	LOW	Row Total
High	12999	1414	14413
	0.568	0.062	
LOW	6766	1696	8462
	0.296	0.074	
Column Total	19765	3110	22875

With the model we have built, in absence of injury nature category, the model was able to classify more generally through the other variables. we were able to achieve 65% of accuracy.

Key findings from the plot:

- Root node of our decision tree was split using body part region , into 3 categories namely {neck , Trunk}, { Lower }, { Head, Multiple and other}
- If injury occurs in neck or trunk the claims are more likely to be high , out of 11533, 9311 tends to be higher risk claims
- When injury occurs in the head or if it is a multiple body part injury or other regions, then claims are more likely to be dependent on the non-standard code mentioned, then those will be low risk claims. Unfortunately if it region is not available those are considered to be high risk claims, without any further considerations.
- If it occurs in the head and if the person is female and days to close the claim takes more than 616 days, then it also results as high risk claims. Out of 1222, 734 claims found to be high risk one.
- The next major splits that occur throughout the plot is days to closure. If it is more than 6245 days majority of the 70% claims turns out to be High risk claims.
- If it is taking more number of days to process the claim and if it turns out indemnity or medical it also turns out to be high risk claims.
- When the number of days to process the claim are less than 625, then upper wage limit turns out to be a crucial factor. If upper wage is less than $1000, then majority of the claims turns out to be higher risk one.

Apart from this there are many other findings, that doesn't have huge impact, but still considerable one which can be derived from the plot.

Recommendations and Analytics Plan

Recommendations

- From our initial exploratory data analysis, Strain and Contusion amounts for a 54% percent (refer support_1 viz) of the claims which also drives the major part of the costs. Since we do not know the nature of work that is causing these injuries so the company needs to investigate at ground level on what type of work is actually causing these injuries and if necessary come up with workshops or improve their machinery to mitigate these injuries.
- Strain and contusion occurs almost every day of the week however there are other injuries like freezing, burns which occurs only on specific week days. In order to reduce their overall costs while implementing any action plans at ground level, the company should implement plans considering the weekday into account instead of implementing plans throughout the week for certain injuries.
- From the decision tree plot of the C4.5 model, we found that majority of the injuries that occur in neck or trunk body part region turns out to be high risk claims. So this is the other injury nature that the company needs to look into more depth.
- Number of days to process a claim is directly proportional to the costs involved. In particular for certain injuries if it takes more days to process, the claim is highly likely to be of high risk one so they should reduce the number of processing days for certain injuries to minimal.

Analytics Plan

In order for the claims company to become an analytics 3.0, the company should go through these changes -

- The company should aim to collect more data from varied sources regarding their employees. For example, employees past work experience details can be taken into account.
- The company should afford to appoint a chief analytics officer and build a dedicated analytics team to monitor the data pertaining to employees. This will help them to flag the high cost claims early and expedite those cases or avoid similar cases using past history.
- The company should try to collect health history of its employees and apply analytics on it. An analytics product where they could come up with a meal plan for its employees based on their health history and the nutrients the employees are lacking which might be making them vulnerable to certain injuries like Strain and Sprain.