BIG DATA ANALYTICS FOR WORKERS' COMPENSATION CLAIMS PROCESS

By: Team Number 4

Hozefa Haveliwaala
Devarsh Jhaveri
Veda Samhitha Vadlamudi

Executive Summary

The problem background and data exploration of the claims dataset were completed earlier and this report is focused on modeling and recommendations. The cleaned claims dataset is merged with the transactions dataset. The newly merged dataset is again then cleaned using SAS Enterprise Guide and excel. Two new derived independent variables are taken from the merged dataset which are 'new_weekday' and 'date_closed-date_received'. Also, another binary dependent variable called 'new_ishighrisk' is added to the dataset.

The dataset is now explored in Tableau and the additional insights which are obtained after adding the three new variables to the merged and cleaned dataset are presented. These visualizations include the derived and binary variables.

Based on the various predictive models and insights from the visualization analysis, the variable 'isDenied' is considered for building predictive models on the merged dataset. The predictive modeling techniques logistic regression, naïve-bayes classification, random forest and Classification and Regression Tree (CART) are compared and the independent variable 'isDenied' is chosen for each model from the list of available variables. After considering the limitations of each of the predictive modeling techniques as discussed in the Appendix, logistic regression is considered as the best predictive modeling technique for this dataset.

In logistic regression, isDenied is considered as the target variable. The data is then partitioned into two segments of which 70% of the data is used for training and the remaining 30% of the data is used for testing. In the training dataset, the input variables whose ChiSq value is greater than 0.5 are not considered for modeling. The remaining variables are considered for modeling. With these remaining variables, a logistic regression model is obtained

Various conclusions are drawn from the results about the impacts of the independent variables on the chosen outcome variable 'isDenied' such as the odds of isDenied increasing with the bill value, odds of isDenied increasing when isHighRisk=0 and the odds of getting denied decreases if the age is more than 70.

A strategic set of recommendations for the claims processing organization are provided which are justified by our data analysis as shown in the appendix.

INDEX

Appendix

1.1 New variables created

We have created the following derived variables from the initial datasets to improve our analysis:

1. new_Weekday
2. date_closed-date_received
3. new_IsHighRisk

Description of the variables:

1. new_Weekday: -
   This variable denotes the weekday on which the incident has occurred. For example, Monday, Tuesday etc., We have used a numeric denotation to indicate the weekday for the convenience of computation. Weekday has been chosen as the new derived variable because it helps in finding out the frequency of any injury which takes place on daily basis or occurs periodically and helps us find out the reason or issue leading to the injury on those days and propose plans to eradicate the cause of injury.

2. date_closed-date_received: -
   This is the second derived variable chosen. This is the date difference between the date when the claim has started processing and the date when the claim was closed. This variable was chosen to find out the duration of time taken by the insurance company to process claims for various types of injuries. This variable gives us another insight about reopened claims and helps us find out about the frequency of claims being reopened and the nature of claims that are reopened. It was noticed that there was a great difference between these two dates for certain types of claims for which the relation and reason is unclear.

3. New_IsHighRisk: -
   This is a categorical binary variable that indicates if a claim falls into the high-risk category or not. This variable is calculated combining the variables: Total paid for body part region, Claimant type and Age.
   i) For a specific body part region, if the value paid is above 90^{th} percentile, the weightage is given as 2, and if it is between the 75^{th} and 90^{th} percentile, the weightage is given as 1 and in all other cases the weightage is 0.
   ii) For the claimant type, if there is a claim for indemnity the weightage is 2, if the claim is only for medical expenses the weightage is 1 and for report it is 0.

iii) For age, if the person is from the age group 40-50 the weightage is 1. Else it is 0.

Using this weightage criteria, a total weight is calculated by adding all the weightages allocated to each variable in each claim case. If the total weightage is equal and above 4, then it is deemed as a high-risk claim and if the total weightage is less than 4 it is deemed as low risk claim.

1.2 Analysis of Predictive Modelling Techniques: -

1.2.1 Predictive Modelling Techniques Compared: -

Predictive Modelling: - Predictive modeling uses statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place.

The initial four predictive models taken under consideration are: -

a) Logistic Regression
b) Naïve Bayes Classifier
c) Random Forest Regression
d) Classification and Regression Trees (CART)

a) **Logistic Regression**: - Logistic regression is a technique in which unknown values of a discrete variable are predicted based on known values of one or more continuous and/or discrete variables. Logistic regression differs from ordinary least squares (OLS) regression in that the dependent variable is binary in nature. This procedure has many applications. In biostatistics, the researcher may be interested in trying to model the probability of a patient being diagnosed with a certain type of cancer based on knowing, say, the incidence of that cancer in his or her family. In business, the marketer may be interested in modelling the probability of an individual purchasing a product based on the price of that product. Both of these are examples of a simple, binary logistic regression model. The model is "simple" in that each has only one independent, or predictor, variable, and it is "binary" in that the dependent variable can take on only one of two values: cancer or no cancer, and purchase or does not purchase.

b) **Navies Bayes**: - Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

c) **Random Forests**: - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

d) **Classification and Regression Tree (CART**): - CART also Known as Decision Tree. Decision tree learning uses a decision tree as a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning.

Comparison between the above models: -

Serial No.	Model	Output Variable	Input Variables	Pros	Cons
1	Logistic Regression	Is Denied	Age, Claimant Type, Is High Risk, BillReviewALE, Hospital, PhysicianOutpatient, Rx, Injury Nature, Body Part Region	Output variable is binary, we get the general probability of the output variable, multicollinearity is not an issue	Does not perform well for large number of categorical value,
2	Naïve Bayes	Is High Risk	Age, Claimant Type, BillReviewALE, Hospital, PhysicianOutpatient, Rx, Injury Nature, Body Part Region	Easy to implement	Relies on independence assumption

3	Random Forest	Is High Risk	Age, Claimant Type, Is High Risk, BillReviewALE, Hospital, PhysicianOutpatient, Rx, Injury Nature, Body Part Region	Handle large number of categorical value	Requires a large amount of memory
4	CART	Is Denied	Age, Claimant Type, BillReviewALE, Hospital, PhysicianOutpatient, Rx, Injury Nature, Body Part Region	Handle large number of categorical value, Output can be visualized clearly,	Requires a large amount of memory

Table No 1: Differences between different types of predictive Models

1.2.2 Details of Predictive Modelling Technique Chosen: -

From the above four models, we have chosen Logistic Regression as the most appropriate model for this dataset.

1) The target variable that would be appropriate for building predictive models on the merge dataset is IsDenied.
2) We have partitioned the data set into two sets i.e. training set and validation set in the ration of 7:3, so that the unused data could be helpful in testing the results of our modeling technique. For this task we have used the SAS Enterprise Miner.

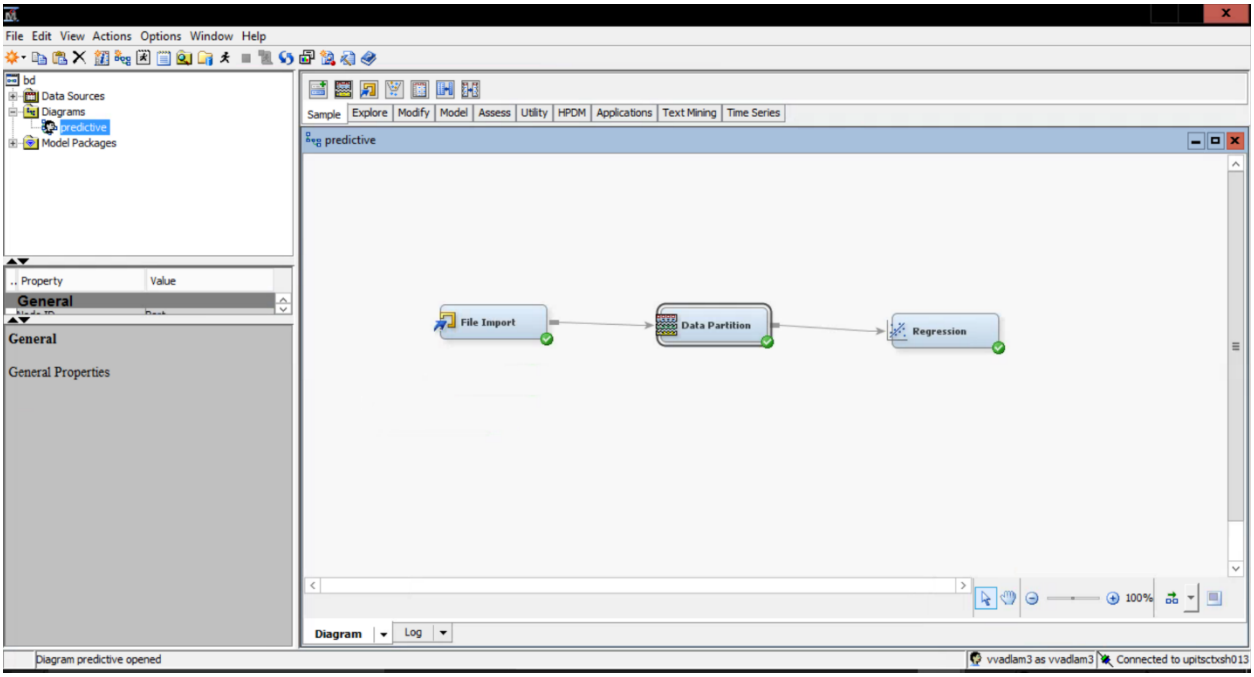


Figure 1: SAS Eminer Process view of data Partition

3) As the second step of the process we considered the following variables as input variables
   a) BillReviewALE
   b) BodyPartRegion
   c) CalaimantAge_DOI
   d) ClaimantType
   e) Hospital
   f) InjurtNature
   g) PhysicianOutPatient
   h) Rx
   i) TotalPaid_End
   j) TotalRecover_End
   k) TotalReserves_End
   l) New_ISHighRisk

4) After running multiple iterations, we have noticed that only the below 6 variables (fig 2) have significant ChiSq value, hence only these are considered for modelling.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
BillReviewALE	1	5.2589	0.0218
BodyPartRegion	6	60142.0356	<.0001
ClaimantAge_at_DOI	70	3661.9572	<.0001
Hospital	1	20.9837	<.0001
PhysicianOutpatient	1	112.2468	<.0001
new_IsHighRisk	1	232.1333	<.0001

Figure 2: Analysis of Effects

4. The following outputs have been obtained: -

Classification Table

Data Role=TRAIN Target Variable=IsDenied Target Label=IsDenied

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	94.764	99.998	40848	94.7618
1	0	5.236	100.000	2257	5.2359
0	1	100.000	0.002	1	0.0023

Data Role=VALIDATE Target Variable=IsDenied Target Label=IsDenied

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	94.763	99.997	30636	94.7603
1	0	5.237	100.000	1693	5.2366
0	1	100.000	0.003	1	0.0031

Figure 3: Classification Table for TRAIN data and VALIDATE data

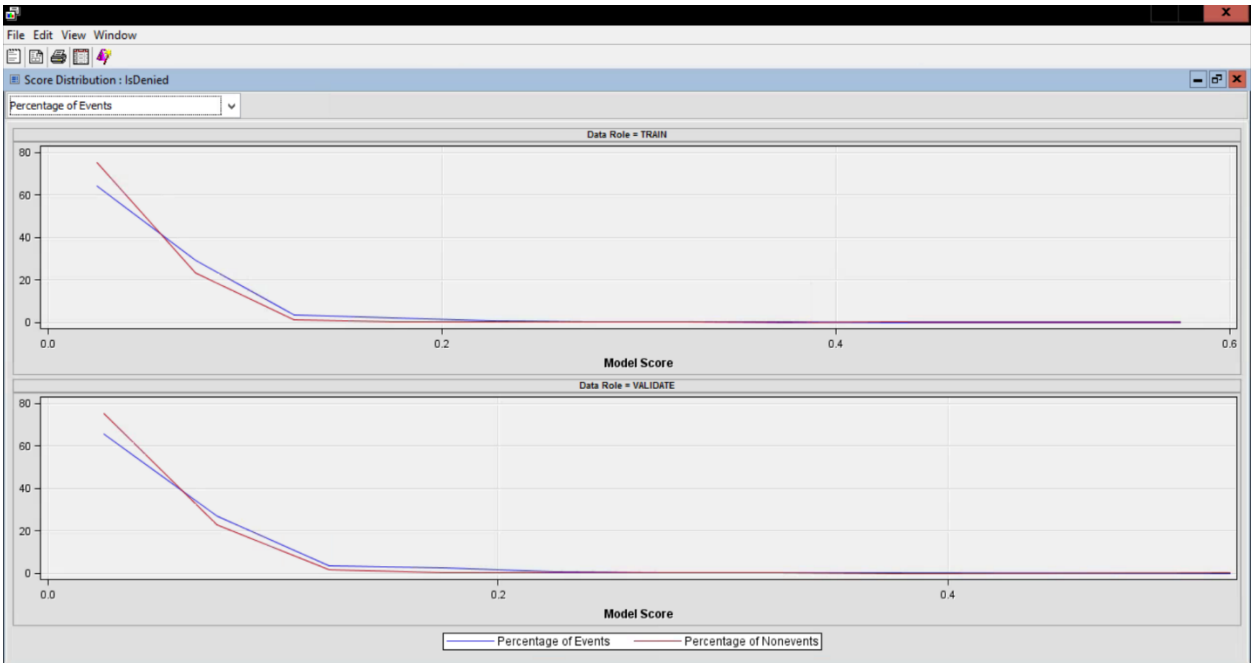


Figure 4: Score Distribution: IsDenied

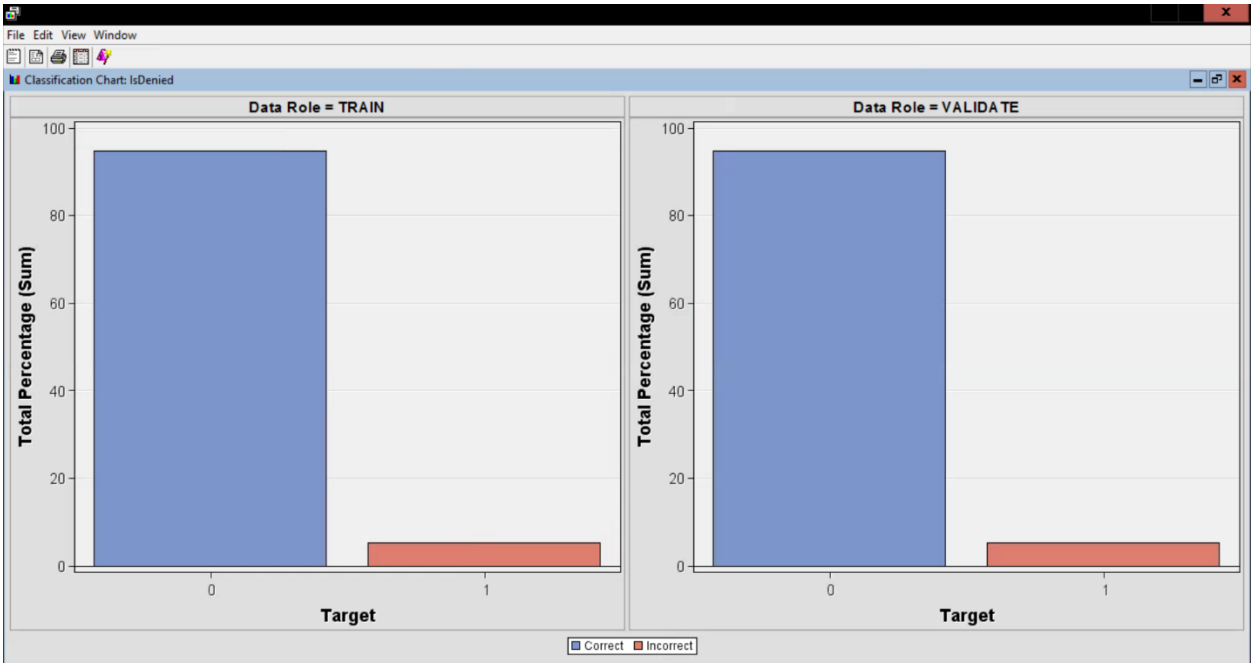
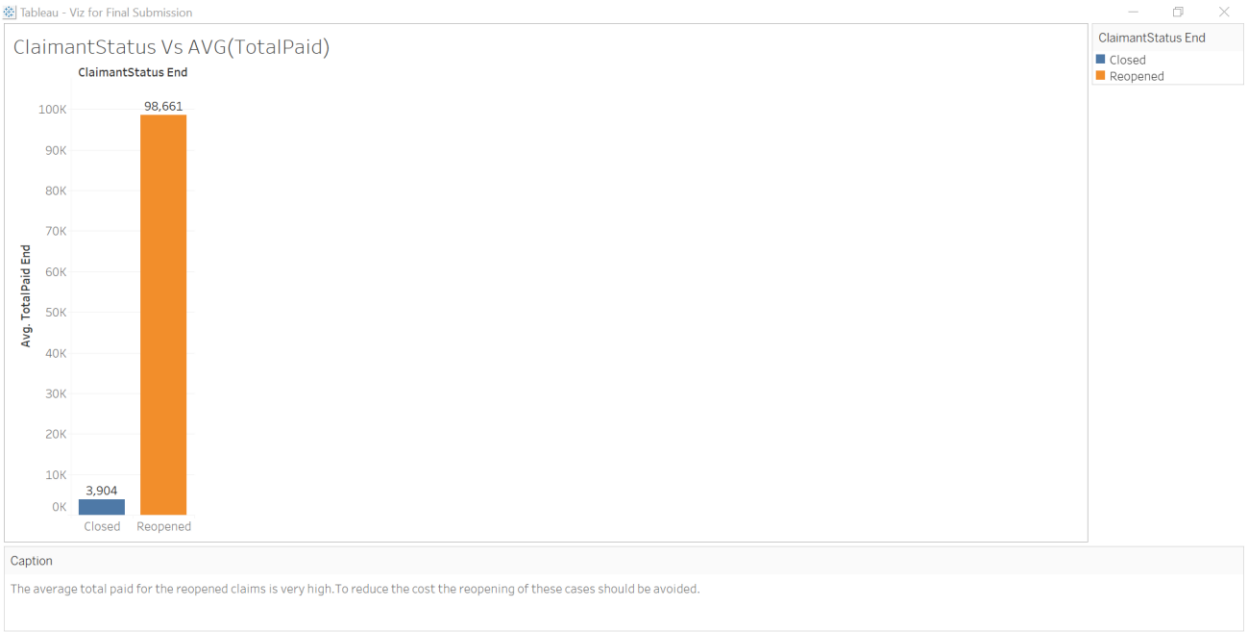


Figure 5: Classification Chart: Isdenied

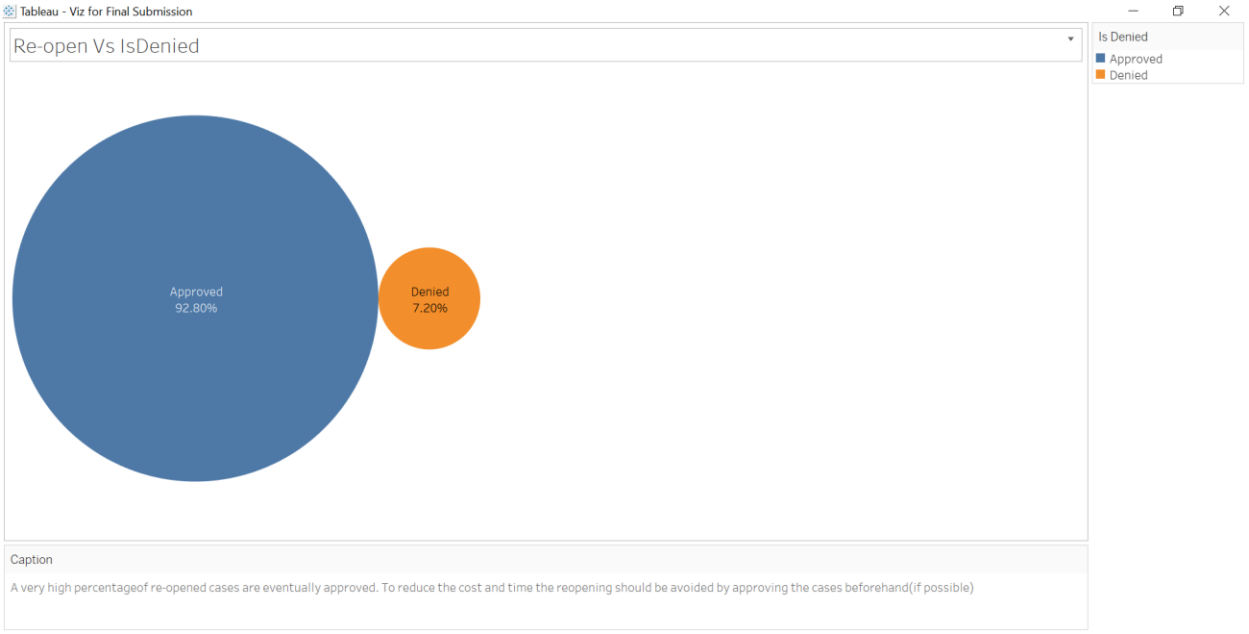
5. Results from the modeling are: -
   a. As the Bill value (Hospital, out-patient) increases the odds of IsDenied also increases.
   b. According to our calculated variable new_IsHighRisk , if it is a high risk claim i.e. new_IsHighRisk is 0 the odds of claim getting denied decreases.
   c. The odds of getting denied decreases if the age is more than 70.

1.3 Recommendations with Justifications

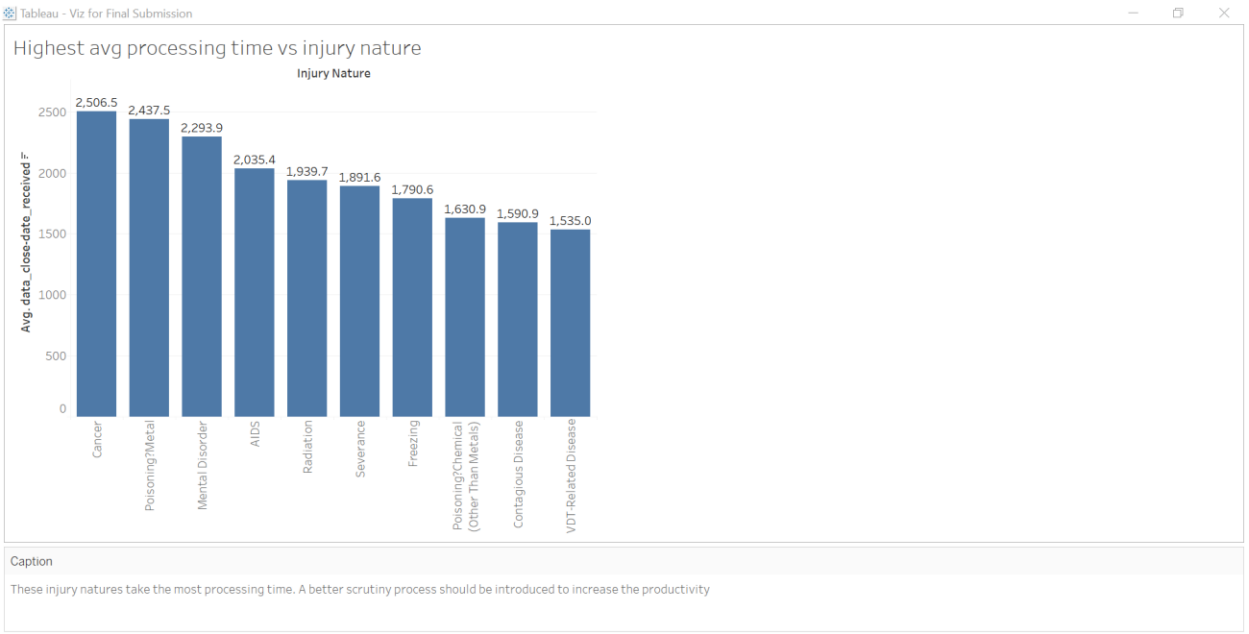
1. High Value claims should be scrutinized more to prevent reopening (if any) to reduce the cost and time.



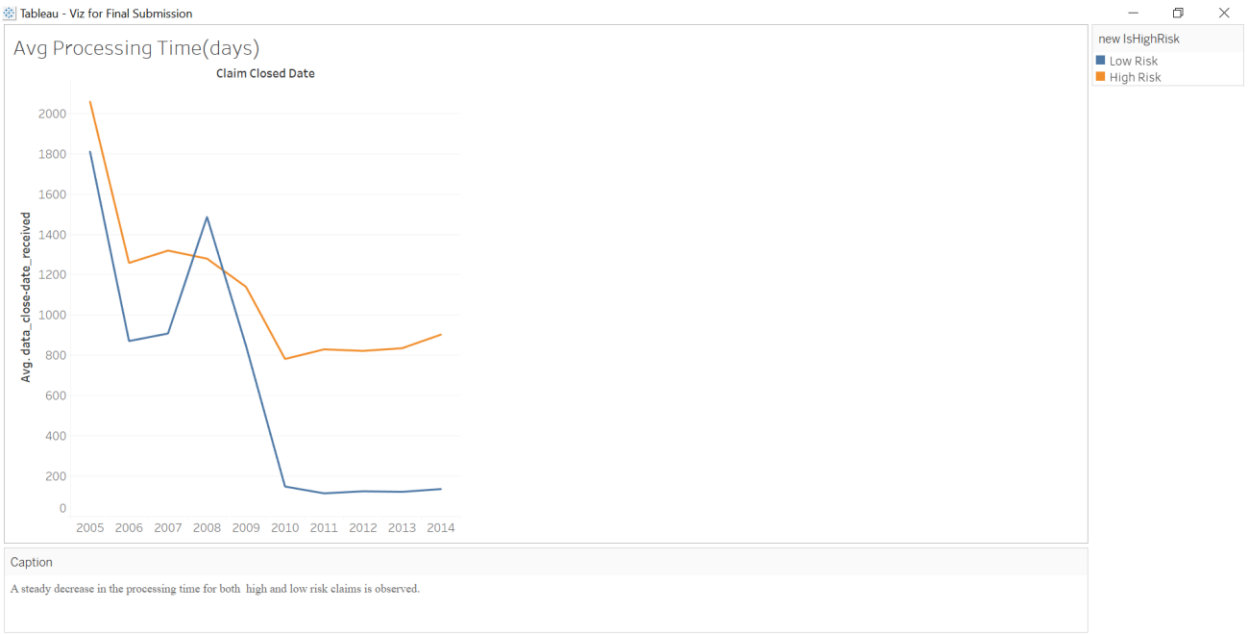
2. Considerable amount of reopened cases is eventually approved. This should be avoided to reduce the additional time and cost.



3. Few Injury natures like Cancer, Poisoning metal and mental disorder takes the most processing time. Amends should be made to reduce this processing time.



4. The average processing time for claim processing has decreased considerably over the years. This should be maintained consistently for further productivity.



1.4 The Analytics Plan

The above organization can become analytics 3.0 competitor by implementing the following key takeaways: -

1. Gathering the external data to improve the analysis: - Firstly the Insurance Company should collect data from various data sources. The data currently present is insufficient to make the required conclusions and understand the trend. There should be many more data collecting points and much more variety of data collected. For example, data about the changes in the Government Policy during the years which effected the rate of acceptance of the claims from the insurance company, or the details of the department in which the employer works which would help us if the injury was related to the department in which the worker was working or the nature of work, etc.

2. Employ people with cross disciplinary knowledge/ experience: - Cross disciplinary team members are always a benefit to the analysis process they help us understand the details of their specialization better which would make it clear to know which variables from the data can be related and watched together for dependable behavior. For example, a person with knowledge about the injury nature would be a great asset in understanding the reason of injuries in a age group, to help the insurance company judge if the claim amount requested is worth the injury nature and the future risk associated with the injury treatment which would lead to reopening of the claim.

3. Use prescriptive analysis to get recommendations to optimize the productivity: - Prescriptive models involve large-scale testing and optimization and are a means of embedding analytics into key processes and employee behaviors. They provide a high level of operational benefits but require high-quality planning and execution in return. Prescriptive analysis helps us find the relation between various variables in the dataset. This in return would make it easy for the analysis of the data collected in future as we already know the relation between them we can predict the nature of output of few variables once we know one of the related terms. Prescriptive analysis tells the insurance company on how it should strategize its process in future and