**Report on Claims Analysis**

This report concentrates on understanding various predictive models i.e. decision trees, linear regression and logistic regression for the claims dataset to understand the influence of various independent variables identified on total payouts and processing time. An initial hypothesis is generated and we run the model to verify the hypothesis. The results are analyzed and recommendations are made to the claims company on how can they can improve on grounds of total payouts/processing time.

**Part 1:**

Initial hypothesis

The easiest relationship to look at would be the relationship between the payout received and the three identification variables: gender, age and body part region (or body part) injured. This regression should be run first to see if there are any obvious determinants to injuries with high payouts. For example, do women or younger individuals receive higher payouts for similar injuries? Perhaps it may also be useful to add the injury nature but this may not be viable due to the large number of different injuries (burn, laceration, strain, foreign body, etc.) presented in the dataset. As with most regressions, this model likely exhibits some degree of collinearity (for example, men generally are more likely to work in positions involving manual labor and thus are more likely to suffer from back injuries and muscle strains than women). However, it seems unlikely that any correlation between any two variables would be strong enough to corrupt the results. We are already aware from the first report that some body injuries receive higher payouts (for example head and neck injuries pay more than injuries to the extremities). Our assumption would be that, adjusting for region of injury, younger claimants will receive higher payments, as they have lost the greatest earnings potential as result of their injury.

The dataset of claims injuries is massive, and the majority of entries contain more or less routine workplace injuries. We previously determined that by far the most common injuries represented in the sample were muscle strains. Perhaps we can conclude that a certain number of bruises, cuts and strains are more or less inevitable for a large operation. These do not earn large compensation payouts, and would take a substantial amount of oversight to even slightly reduce. For this reason is may be important to sort injuries payouts into quantiles to only look at “large” payouts. This can be accomplished by creating a binary variable representing “large” payouts. In our case, we decided to label any payout over the mean of $300 as “large.” We can then run a logistic regression using large payouts as our binary dependent variable and the same independent variables as before.

Hypothesis 1:

However, the previous regression would probably not reveal many new insights to the company, as they are likely already aware of what payouts to award certain individuals and injuries. The first regression would simply serve as a benchmark to see if anything unusual was going on. For example, the company may be concerned if the regression results reveal that older individuals receive higher payouts or if finger injuries received higher payouts on average then, say, head injuries, as those findings would not make a great deal of logical sense. However, our first report did find something unusual was occurring in the year 2009. Namely, a great deal of both claims and payouts were made. There are two reasonable explanations for this phenomenon. First, possibly the company changed their workplace environment in a way that made workplace injuries, and thus claims and payouts, more likely. While possible, we do not really have a way of testing this hypothesis. Another possible explanation is that the company changed the way claims were filed and payouts were awarded in a way that caused both of these factors to increase. We can investigate this issue by regressing the total payout received with our three determinants (gender, age, injury body part/region, possibly with the addition of injury type) and a binary variable representing claims opened in the year 2009. If there was no policy change, than we should see no effect from the addition of the binary variable. It may also be viable to create a binary variable for claims opened after 2009 (since the number of claims made decreased to more “normal” levels after the 2009 spike) and rerun the regression. If a policy change did not occur, the post-2009 coefficient would be zero. Because there is a sizable gap between the when a case is filed and when a payout is made, it may be useful to perform the same regressions above only now using payouts made in or after 2009 as an independent binary variable instead of claims opened. Note that it would not be a good idea to include both claim opened and payout made binary variables in the same regression, as it would likely result in a large amount of multicollinearity.

Hypothesis 2 :

Another potential approach would be to look and see whether claims filed at certain times of the year or certain days of the week are more successful than others. To test this, we would need to create eleven month and six day of the week binary independent variables (with one omitted from each to avoid perfect collinearity). These would be regressed along with injury type, gender, age and whether the claimant was fatally injured (assuming fatalities lead to higher payouts) on payout received (our dependent variables). Our hypothesis would be that payouts are agnostic to when they are filed and are determined by the nature of the injury and the nature of the claimant.

Hypothesis 3:

The time gap between when a claim is filed and when the payout is made is also a topic of interest. One hypothesis is that claims with higher payouts will take longer to process. This can be tested by creating a variable for how many days have passed between the day a claim is filed and the day a payout is received (note that in the dataset that claimants who receive no payout still have a payment date listed), then regressing payout received with the independent variables listed above and days between filing a claim and receiving a payout. If our hypothesis is accurate, there should be a positive relationship between the number of days and the size of the payout. However there is also a chance that claims that take more time to process are investigated and vetted more thoroughly and are thus associated with lower payouts.

Hypothesis 4:

Another time difference that may be important to observe is the number of days between the day the incident happened and the day a claim was opened. In theory, if a worker takes a longer time to open a claim, the company and insurer would be more likely to disprove the worker’s claim and award a smaller payout or no payout at all. However, much like the previous paragraph, the relationship between payout and time between the incident date and opening date is not an obvious one; there is a chance that complex cases involving severe injuries (which would be associated with high payouts) take time to open or cases opened the same day as the incident occurred may be more likely to be investigated as fraudulent and award no payout. This can be tested by regressing the size of the payout awarded with our previous control variables age, gender, body part/region and the gap between the incident date and the date the claim was opened. We could also run a logistic relationship using the same independent variables and the the dependent variable “is denied.” In this case, our hypothesis would be that a larger number of days makes a worker more likely to be denied a claim.

Hypothesis 5:

Another variable to examine is the processing time. This variable may have the same determinants as payout size. For example, a fatal injury may warrant a high payment, but could also take a considerable amount of time to process. This can be tested by running a linear regression using processing time (the number of days between the date a claim was opened and the date a claim closed) as the dependent variable and the previously listed independent variables age, gender, region of injury and whether the injury was a fatality. Like with payouts hypothesis listed above, we can also group processing times into quantiles to only focus on claims that took a long time to process. In our case, we created a binary variable “Processing Time Binary” representing a “long” processing time, denoted by having a processing time greater than 700 days. We can then run a logistic regression using that binary dependent variable and the independent variables listed previously.

**Part 4:**

Linear Regression:

For a linear regression, it would be useful to determine if claims filed on certain months or days of the week were more successful than others. To do this, we will have to create a set of binary variables representing each day of the week and each month of the year. These will serve as some of our independent variables. The dependent variable would be total payout received (named “TotalPaid” in our dataset). In total, we should use eleven month binary variables (omitting one to prevent perfect multicollinearity) and six day of the week dummy variable (again, omitting one to prevent perfect multicollinearity) as independent variables in this regression.

We also need some other independent variables to control for other circumstances. First, we will include a binary variable that takes a value of one if the claimant is male. Generally, since men tend to work in more labor-intensive positions than women and because more claimants are male, the coefficient related to this term should be positive. Next, we need to include the body part region injured. Again this is accomplished by creating seven binary variables representing each body region (head, neck, upper extremities, lower extremities, trunk, multiple parts and non-standard code). Six of them would be used as independent variables (to prevent perfect multicollinearity). It may be heard to tease any exact relationships between injury region and payout, but we should expect to see head and neck injuries, as they are usually more severe, to see higher payouts than the others. Cases with multiple body parts injured could also pe positively correlated with payouts. It is important to note that we are using the body part *region* and not body part as our independent variable because there were simply too many body parts and combinations of body parts listed to use them all in a single model. We did not include the type of injury suffered for the same reason. This is a major weakness of the model. Choosing body part region may be too broad, and will not be a good indicator of the severity of the injury (for example, a skull fracture and a nose bleed are both classified as a head injury in the dataset).

We should also include the claimant age as an independent variable in this model. We would assume the coefficient related to this variable would be negative. As a claimant ages, their earnings potential declines, and thus, they have less to lose than a younger worker in the event of an accident and may receive lower compensation as a result. In addition, we will include the binary variable representing whether or not a claimant is fatally injured. We should expect this variable to be positively associated with payouts. Finally, we will include the processing time as an independent variable. Perhaps surprisingly, the collinearity between processing time and payouts is rather small (0.15). For this reason we can use processing time as an independent variable without having to worry too much about multicollinearity.

Logistic Regression:

Another approach is to use a logistic regression to understand determinants of critical payouts. One can argue that a certain amount of workplace injuries are minor in nature, such as cuts, bruises and minor strains. These injuries result in few hours or days of work lost and compensation claims will result in very small payouts. Preventing these injuries would be more or less impossible. Therefore, it may not be viable to look at these “inevitable” injuries and instead look at preventable injuries that are severe in nature and often lead to very large payouts. These will cause the greatest loss from the company’s perspective and in theory could be prevented. The dependent variable in this case would be a binary variable representing a major injury. In our case, “major” injuries are injuries that receive payouts greater than $300 (approximately the mean payout). This variable is labeled “Total Paid Binary.” Note that we are assuming severe injuries receive higher payouts, which is not a ground-breaking assumption.

The independent variables are similar to those described in the linear regression. Again, we can use six day of week and eleven month of the year the report was filed as control variables (once again omitting one for each to prevent perfect multicollinearity). The initial assumption would be that these results should be insignificant, as there should be biases between when a report is filed (which can be several days after the incident actually happened) and the payout. We will also include binary variables representing male claimants and the body part region injured (head, neck, upper extremities, lower extremities, trunk, multiple parts and non-standard code, omitting one to prevent perfect multicollinearity). Another independent variable used will be the age of the claimant, as will a binary variable representing a fatal injury. Finally, we can use the processing time as an independent variable. Again, the collinearity between the Total Paid Binary variable and processing time is surprisingly very small (0.021), allowing us to use processing time as an independent variable without concern for multicollinearity.

The major drawback of this model involves the classification of the Total Paid Binary variable. $300 is still a fairly small compensation, and may also include many “inevitable” injuries. However, increasing the critical point brings forth the possibility that preventable injuries will not be included. Our assumption is that $300 payouts are high enough to include the majority of preventable injuries (and the preventable injuries with payouts below this may not award payouts large enough to invest in correcting) while omitting most inevitable minor injuries. If this assumption is not the case, we could have a problem. However there is no easy way to test this assumption without manually reading each case and determining if it is severe and preventable or not. Perhaps we could examine the type of injury and sort specific types of injuries as severe and preventable or not, but this would not be entirely accurate. For example, a burn could represent a small burn to the hand (perhaps an “inevitable” injury) or full-body burns caused by an explosion (severe and probably preventable).

Decision Tree:

A decision tree helps us understand the the factors contributing in driving the consequences. Since we are looking at looking at Claim payments and Claim processing time, we will have to select variables that might influence these decisions more acutely.

For Claim payment decision we would look at the following variables:

* Dependent: Total Paid Binary. It is easier to understand a yes or no from a binary variable as an end decision.
* Independent:

1. Gender: We discovered from our earlier exploratory that women delayed reporting any injury. Therefore, this could be a very insightful variable as it encapsulates time factor too while making the claim.
2. Injury Nature: Certain injury nature are decisive if the payment is to be made or not. It is often accompanied with other variables to make a valid call.
3. ClaimantType: There are 3 types of claims in the given dataset- Medical, indemnity and report only. It is possible that at time the indemnity is not needed to be paid. Also, report only type of claims are not paid. this could create a clear bifurcation in the decision tree.
4. InjuryNature: There have been instances in data where the claimant was not paid for severe injuries. This could tell us more if injury type is in fact a deciding factor
5. IsDenied: It is interesting to see that Total paid binary is sometimes 1 even when the claim is denied.
6. BodyPart: We want to see is if Body parts play a role in overall decision making. For example a head injury is considered more severe than a wrist injury. Does such difference in severity affect?
7. BodyPartRegion: Since we are taking BodyPart into account, it is only logical to drill down further to check if a particular region can drive the decision.

For Claim payment decision we would look at the following variables:

* Dependent variable: Processing Time Binary
* Independent variable:

1. Payment\_Amount: We want to see if a certain threshold on payment amount takes more time to process. It may happen to be that larger payouts take more time because they need to be investigated and verified.
2. Gender: As we have discovered that women report an injury. It is a known fact that the later you report an incident to the insurance company, the lower are the chances of processing and the longer the processing time. Therefore, we believe Gender can be a very crucial factor.
3. Count\_of\_TransactionID: As the number of transactions increase, the Processing time binary variable tend to be 0.
4. IsFatality: We want to see if something instantly identified as fatal is process faster or slower. Or the injuries which are not fatal cost tend to cost lesser and is so processed faster?
5. ClaimStatus: We noticed that the processing time binary for Claim status which was reopened is 0 for all. That means that they take relatively lesser time to process. Therefore, this variable too can create a clean classification for processing time.

The result from this model will give us the crucial and decisive variables in arriving to a decision whether or not the claimant is to be paid and in what time is he likely to be paid. We can do this by looking at the historical data and building the decision tree based on that and then applies the algorithm on the newly applying claimants. This can make the task of decision makers much easier and it would no longer be on the intuition whether or not a claimant is to be paid and what is the approximate wait time. It can be based on previous learning and a proper algorithm.

Business problems of different models used:

**Logistic regression** predicts outcomes based on selected independent variables. These variables are selected by the analyst or data scientist. Therefore, the analyst/scientist needs to have a good understanding of the business too to be able to select vital variables. But if researchers include the wrong independent variables, the model will have little to no predictive value. Logistic regression works well for predicting categorical outcomes. It can also predict multinomial outcomes. But the biggest drawback is logistic regression cannot predict continuous outcomes. Therefore, we cannot solely rely on Logistic regression for our predictions. The next biggest danger with Logit is it is vulnerable to overconfidence with its accuracy. The model can appear to have more predictive power than they actually do as a result of sampling bias. It is highly likely that it overstates the accuracy of its predictions.

**Linear Regression:** It is often not necessary to run linear regression on a nonlinear relationship. It is only after the model is tried to fit on the data that we realize it has failed. It leads to a lot of waste of time and money. The next drawback of linear regression is that it can only run numerical values. If the data does not fit the model the deviance is very high and there is nothing that can be learnt from the result.

**Decision trees** function on the input fed at the very beginning of the tree. A minute change in the input values can give an erroneous decision in the end. Preparing a decision tree without proper expertise, experience, or knowledge can cause garbled outcome of business opportunities or decision possibilities. Also, decision trees are supplemented with excessive information which can lead to “paralysis of analysis” . There are good chances that decision makers might feel burdened with such excess of information and this may lead to slower decision making in the organizations. Decision trees are biased towards choosing the higher values for branching options but it is not necessary that the variables with higher values are more decisive. It does not drop missing values. It adjusts itself to account for the missing values.

**Part 5:**

A linear regression model described in the section above was run on the data. The results of the regression can be found in Table 1. Some of our assumptions hold. For example, fatal injuries are associated with higher payouts. Specifically a fatal injury generally leads to an approximately $1,400 increase in total payouts received. In addition, processing time has a positive and significant relationship with payouts. One additional day processing a claim corresponds to a roughly $4 increase in payout. This seems to prove our assumption that claims with high payouts take more time to process.

However, not all of our assumptions hold true. Interestingly, age is positively correlated with payouts. Older claimants tend to receive higher payouts than younger claimants. The coefficient is rather large too. A one year increase in age is associated with a $137 increase in payout. Male claimants are positively correlated with high payouts. This is likely caused by men generally working more labor-intensive jobs compared to women. However, the magnitude of the effect is startling, On average, male payouts are almost $600 more than female payouts. The body injury region has some interesting results, but the results are not likely meaningful. The lowest payouts are associated with injuries that cannot be classified, which makes sense as these injuries are likely harder to prove. Surprisingly, head injuries receive lower payouts compared to many of the other injury regions. However this is likely caused by the generalizations made by using this variable that were described earlier (such as “head” injuries including everything from nosebleeds to skull fractures). In order from highest payouts to lowest, the body regions are trunk, multiple parts, neck, lower extremities, upper extremities (the omitted variable), head and non-standard form.

The strangest results can be seen when looking at the day of the week and month of the year controls. There are two significant months, April and June, where payouts are extremely high compared to all other months. Claims opened in June on average pay $3,400 more on average. April claims earn almost $1,400 more on average. This does not make sense. Unless major accidents are simply more prone to occurring in April, this must be caused by the claims company over-awarding claimants who filed during these months. Perhaps there is an explanation for this (maybe involving fiscal quarters and how liability insurance is “pooled”),but nevertheless this issue needs to be investigated by the firm.

Even stranger is the day of the week results. The only significant day is Sunday, which is considerably large. Strangely, claims filed on Sunday pay on average almost $3,500 more. This is very bizarre as filings are often made days after the actual injury. There should be little relationship between when a claim is opened and the payout, as there is usually a few days between when an incident happens and when a claim is filed. However, this does not necessarily mean that workers should wait to open their claims on Sunday to receive more money. Likely this is caused by the fact that only a small portion (0.2% of the total sample) of claims are filed on Sunday. However, it may warrant further investigation to see if Sunday openings are particularly major injuries.

**Part 6:**

Based on the results observed and listed out in Part 5, we would propose the following strategic recommendations for the claims management company so that they can better manage their business. As observed, high payouts take more time to process. But this does not need to be the case. The company needs to review their claim payout process and see if there is any unnecessary procedural overhead for larger payouts. Ideally, this does not need to be the case and the company can be more efficient by streamlining their process.

Regarding the observation, head injuries receive lower payouts compared to many of the other injury regions, the company should research why this is so because the expectation is head injuries will cost more in treatment. Could it be that head injury claims or costs are dealt in a more efficient way than others? If this is so, this is an improvement opportunity for company to review the process for other injuries in other body parts.

Another observation made was related to payouts being extremely high in April and June compared to all other months. The company needs to research more on the exact reason behind this scenario and if it's due to customer behavior or higher accidents happening during that time frame, company needs to look at opportunity to better educate or warn customers during these times.

The fact that only 0.2% of claims are filed on Sunday is strange and the company needs to investigate on what is the cause and also, if this behavior is affecting the company adversely or not. The Company needs to review if customers are having difficulty filing claim on Sundays and if so, the company must try to make their claim filing process for e.g. an online filing system more user friendly or easily accessible or market it better.

**Appendix**

**Details of New variables created**

-> Payment\_Amount = Total Payment Amount has been calculated by aggregating the individual transaction payment amounts for each claim. This was done as part of merging claims and transaction data file. The SAS Enterprise Guide tool was used to perform the required calculations.

-> Count\_of\_transactions = Count of Total Number of Transaction IDs for each unique claim. This calculation was done as part of merging claims and transaction file. The SAS Enterprise Guide tool was used to perform the required calculations.

-> Total\_incurred\_cost = TotalReserves + Indemnity + OtherPaid - TotalRecovery. This was a new column defined as part of Phase 1 of the project and the column was defined and added to the existing dataset using SAS Enterprise Guide.

-> Processing time(in days) = Used SAS Enterprise guide to calculate the difference between claim opened date and claim closed date in days.

INTCK('DAY',t1.ClaimantOpenedDate,t1.ClaimantClosedDate)

**Dependent Variables**

-> Processing\_time\_binary - The cutoff to create the binary variable i.e. critical as 1 and non-critical as 0 based on processing time is 700 which is around average of the processing time for the whole dataset. The number of 0’s and 1’s were almost equally balanced.

Number of Critical(1) i.e. above 700 - 30387

Number of Noncritical(0) i.e. below 700 - 53111

-> Total\_paid\_binary - We didn’t choose the average here because the number of 0’s and 1’s had a lot of difference. To overcome this, we chose the cutoff to create the binary variable i.e. critical as 1 and non-critical as 0 as 300.

Number of Critical(1) i.e. above 300 - 40354

Number of Noncritical(0) i.e. below 300 - 43144

**Independent Variables**

-> Claims\_MonthOfYear = The month in the calendar year on which the claim was opened.

-> Claims\_DayOfWeek = The day of the week on which the claim was opened.

-> GenderF = Recoded column for Gender created using SAS Enterprise Guide. The variable has the value 1 when Gender is Female and has the value 0 for all other Gender types.

-> GenderM = Recoded column for Gender created using SAS Enterprise Guide. The variable has the value 1 when Gender is Male and has the value 0 for all other Gender types.

-> BodyPartRegion\_Head = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Head and has the value 0 for all other BodyPartRegion types.

-> BodyPartRegion\_LowerExtremities = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Lower Extremities and has the value 0 for all other BodyPartRegion types.

-> BodyPartRegion\_MultipleParts = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Multiple Parts and has the value 0 for all other BodyPartRegion types.

-> BodyPartRegion\_Neck = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Neck and has the value 0 for all other BodyPartRegion types.

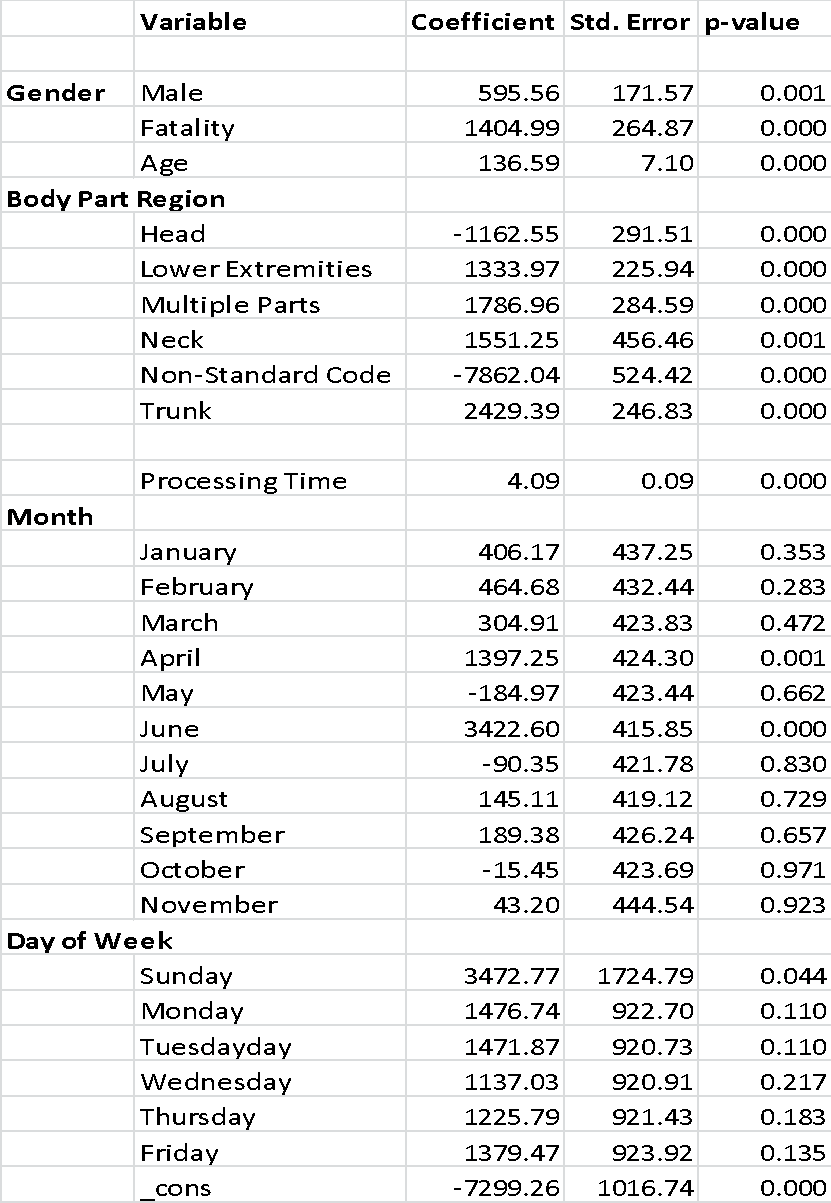
-> BodyPartRegion\_NonStdCode = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Non Standard Code and has the value 0 for all other BodyPartRegion types.

-> BodyPartRegion\_Trunk = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Trunk and has the value 0 for all other BodyPartRegion types.

-> BodyPartRegion\_UpperExtremities = Recoded column for BodyPartRegion created using SAS Enterprise Guide. The variable has the value 1 when BodyPartRegion is Upper Extremities and has the value 0 for all other BodyPartRegion types.

**Output of linear regression**

Table 1: Linear Regression Results



Note: R-squared: 0.0354