Analytics Modeling on Claim Datasets

DSBA 6100 Project Phase 2

Group 2: Xinhui Cai, Micajah Jones, Menghang Li, Shuhua Yin

Executive Summary

Xinhui Cai, Micajah Jones, Menghang Li, Shuhua (Jessica) Yin

The objective for this analysis is to use prediction model to examine our hypothesis derived from phase 1 and obtain some insights. During the process, we examined how each model can be used in this case, evaluating individual strengths and weaknesses along the way. Then we implemented the Logistic regression model to present the results. Based on a thorough understanding, we found major drivers that impact claims processing time and claims payment, along with some strategic recommendations, for improving company's productivity and competitive positioning in the domain.

Our initial hypotheses focus on studying independent variables (Body Part Region, Claimant Age, Total Recovery, Processing Time, Claimant Type, Recovery Period, Claimant Open Day on Week, because we speculate them to have significant impacts on the binary outcome variable Collected Payments. We also think these independent variables might have some important relationship with the other binary outcome variable Processing Time: Total Reserve, Body Part Region, Claimant Type, Gender, Recovery Period, Claimant Open Day on Week, Collected Payment, Fatality (of the injury), and Claimant Age.

Based on our initial hypotheses, we decided to use two separate logistic regression models for Processing Time as the outcome variable and Collected Payment as the outcome variable by using R. Within the logistic regression, we used forward selection method for both models. From the results, there are two important attributes or impacting factors which are Body Part Region and Claimant Type for both two outcome variables. For Body Part Region, R system set Head as control group. Therefore, compared to head, all other body part regions lead to increasing probability of processing time/payment being critical and among them, Multiple body parts has the highest possibility to affect the outcome variables. As for Claimant Type, we found out that Indemnity type is the most impactful factor compared to other types for both processing time and payment. Finally, Fatality lead to processing time being critical and the probability increased by 63%. Claim Open Day on Week affects the processing time by about 50% when each one day increased from Monday to Sunday. According to this logic, claims opened on Monday have the highest possibility to make critical payment appear.

The recommendations that we have decided to provide the company with were found using the assumption that the most effective way of improving organizational effectiveness is by reducing unnecessary technological overhead and administrative costs. We recommend increasing capital reserve requirements for claims with certain characteristics in order to proactively allocate resources. We also recommend that the firm expedite claims with certain characteristics so they can be closed promptly, reducing administrative inefficiencies. Finally, we recommend that the client company start expanding the types of data they are collecting in relation to their employee's workplace injuries so that they analytics team can find different insights in the future.

Project Phase 2 Report

Xinhui Cai, Micajah Jones, Menghang Li, Shuhua (Jessica) Yin

Initial Hypotheses

Given the dataset, we want to investigate the relationships of: collected payment & other variables, and processing time other variables. Thus, we developed the following hypotheses regarding to the possible independent variables that could be contributing significant effects to the target variables:

When target variable is Collected Payment:

1. Body Part Region: We want to see which part(s) of the body contribute to critical/non-critical payment, and if they are significant in the model.
2. Claimant Age: A claimant's age could be an important factor for the condition of the collected payment, and we want to explore how this variable could make the impact
3. Total Recovery: If deducted from the sum of Indemnity, Total Reserves, and Other Paid, to derive the Total Incurred Cost, then Total Recovery amount should make the probability of collected payment less critical.
4. Processing Time: the time difference between Claimant Opened Date and Claimant Closed Date, showing how long the claim was processed
5. Claimant Type: Indemnity, Medical, Report Only—which one(s) make a greater impact on the collected payment? Or does any of these types make any significant factors to the target variable?
6. Recovery Period: the time difference between incident date and return to work date—how long/fast an employee recovers from injury. If the recovery period is long, then it could make the collected payment to be critically high, and therefore a positive relationship between the two.
7. Claimant Open Day on Week: We changed the date to days of the week (Monday to Sunday) as numerical values from 1 to 7. We speculate that Monday will have the most claims and collected payments will change drastically for each claim on this day.

When target variable is Processing Time:

1. Total Reserve: we suspect that an increase in the Total Reserve account of a given claim will decrease the probability of Processing Time being critical for that claim. Our rationale behind this hypothesis is that if a claim already has funds in the reserve account, the costs of the claim can be covered immediately rather than waiting for the funds to be transferred from another source.
2. Body Part Region: we speculate positive relationship between all body part regions and the status of the Processing Time.
3. Claimant Type: Indemnity, Medical, or Report Only. We want to see if there might be a significant relationship between a specific Claimant Type and the status of the Processing Time. the processing time should be the shortest—the probability of Processing Time being non-critical is higher. Therefore, Report Only should not be statistically significant for the model,

then we need to explore if Medical Only and/or Indemnity could potentially increase the probability of Processing Time to become critical and be significant for the model.

4. Gender: the gender of an employee may also be an influence on the status of Processing Time—males and females may make the processing times different because of their physical nature in addition to their injuries. The times males and females take to report injuries could also influence their overall processing times. We want to know when Gender is male, if the probability of processing time being critical will increase.

5. Recovery Period: the time difference between incident date and return to work date—how long/fast an employee recovers from injury. The longer an employee takes to recover, more likely the processing time status is to be critical.

6. Claimant Open Day on Week: the same reason when the target variable is Collected Payment. Processing Time will also change more on Monday than other days of the week.

7. Collected Payment (Total Incurred Cost): the likelihood of processing time being either critical or non-critical could also be related to the amount of total incurred cost

8. Is Fatality: Whether an injury is fatal could be significantly affecting the processing time being critical—it could more likely to be critical if the injury is fatal.

9. Claimant Age: A claimant's age could be important for how long the claim will be processed, and we want to explore how this variable will induce the outcome variable.

Model Comparison

Linear Regression

As for any linear regression, we assume that there is a linear relationship between our dependent variables, total incurred cost and processing time, and our independent variables. By examining which independent variables significantly affect the dependent variable. We can directly use the total incurred costs/processing time as the numerical outcome variable. For the input variables, we may choose those that are relevant to our hypotheses, such as age, gender, body part region, fatality, recovery period, claim open day on week and claimant type, etc. For some categorical variables, we should convert them to dummy variables, otherwise, they could not be applied to the linear regression model. Meanwhile, we may ignore some date-like columns, which are nominal variables, and thus are not suitable for linear regression. Moreover, we may ignore "body part" and "injury nature" variables, each contains too many categories that may vastly increase the model's degree of freedom, in turns, increasing the chance of overfitting our model.

The result of linear regression model clearly indicates which independent variables are statistically significant to collected payment/processing time and how many degrees of outcome changes due to those independent variables changes. The limitations of linear regression are:

- No toleration to missing values, meaning the model is highly demanding in data selection, but missing value problem normally exists in the business dataset.

- Only works for numeric variables and requires outcome variable be continuous. In fact, lots of dataset in business domain are categorical, which increases complexity for the model.
- Linear regression is limited by too many assumptions, such as linear relationship assumption, homoscedasticity assumption, no multicollinearity assumption, etc. However, with respect to business problems, many of them are not linear problems but we are not aware of that.

Decision Tree

For decision tree, it uses classification technique to group observations into categories and finds out the best decision rule. This model tolerates missing values, which was the main problem when we faced in wrangling the data in phase 1. It clearly shows all the splits and significant variables that affect the decision, which is more perceptual for viewing the selection procedure than other models are. In this case, we could convert total incurred costs/processing time to categorical types on specific criterions, for instance, we set 12 months as the divider for processing time and $437.775 as the divider for payment. Furthermore, we use the same predictor variables as in linear regression. However, some continuous variables should be divided into categories for splitting, for instance, the age could be binned to <=18, 18-65, >=65. Besides, we could even add up some variables with a great number of missing values, such as "average weekly wage".

The result from decision tree shows the most relevant variables that affect the processing time or payment amount, in addition, it provides us a path to classify groups and form the prediction model. Limitations of decision tree model:

- The overfitting problem, especially in business domain, there's not enough validation dataset for pruning the model.
- Without having coefficient estimations, it can't clearly explain how much changes independent variables would induce in outcome variable, but that is necessary for explanations in business analysis.

Logistic Regression

For logistic regression, it predicts which independent variables significantly impact the outcome and how much they can affect our predictions. In contrast to linear regression, the outcome variable can only be binary, which is coded as 0 or 1.In our case, before we set the target variable, we should convert the processing time and total incurred costs to binary type based on criteria (12 months as divider for processing time, $437.775 as divider for payment), then we set response variable as "processing time is critical" (or "payment is critical") and code it as 1, conversely, "processing time is non-critical" coded as 0.For independent variables, we could use age, gender, body part region, fatality, recovery period, claim open day on week and claimant type into the model, except for those with missing values(like average weekly wage), apparently logistic regression has low toleration to the missing value. For some categorical variables, like body part region, fatality and claimant type and so on, the logistic model will

convert them into dummy variables. However, in order to avoid overfitting problem, we ignore variables with too many categories, such as “body part” and “injure nature”, etc.

Logistic regression predicts the probability of an event occurring based on the past data, also it chooses parameters that maximize the likelihood of observing the data. In this case, the result indicates which predictor variables are significant to the model, and how strong they associate to “processing time/payment is critical”. If we look into the odds ratio, we could see one unit increase in each significant variable will induce how much increases or decreases in odds of processing time/payment is critical. Limitation of logistic regression model:

- It only works for categorical outcome variable, while lots of business problems are related to numeric/ continuous outcomes.
- Based on assumption of linear-relationship, so it’s not suitable to be applied in non-linear business problem.
- It has no toleration to missing values, which normally existed in business dataset.
- It is sensitive to multicollinearity between variables, and may have overfitting issues.

Analytics Modeling – Logistic Regression

We decided to build two separate Logistic regression model for processing time as outcome variable and collected payment as outcome variable to make a thorough analysis for both time efficiency and payment collection efficiency. “Forward Selection” method is used for both models.

1. **Processing Time as outcome variable**

```
> summary(finalmodel)

Call:
glm(formula = TimeBinary ~ TotalIncurredCost + ClaimantType +
    BodyPartRegion + recovery_period + claimant_age + Gender +
    IsFatality + ClaimOpenDateOnWeek, family = binomial, data = claimData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8288  -0.7685  -0.6838   0.9438   2.6457

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.835e-02  6.713e-02  -0.273  0.78455
TotalIncurredCost              8.832e-05  3.233e-06  27.319  < 2e-16 ***
ClaimantTypeMedical Only      -9.541e-01  3.082e-02 -30.956  < 2e-16 ***
ClaimantTypeReport Only       -2.890e-01  4.016e-01  -0.720  0.47168
BodyPartRegionLower Extremities  8.558e-02  4.339e-02   1.972  0.04856 *
BodyPartRegionMultiple Body Parts  7.957e-01  4.910e-02  16.206  < 2e-16 ***
BodyPartRegionNeck            -8.006e-02  6.739e-02  -1.188  0.23486
BodyPartRegionNon-Standard Code  -1.940e+00  3.357e-01  -5.781 7.43e-09 ***
BodyPartRegionTrunk            2.989e-01  4.539e-02   6.587 4.49e-11 ***
BodyPartRegionUpper Extremities  1.336e-01  4.140e-02   3.228  0.00125 **
recovery_period                2.434e-03  2.131e-04  11.419  < 2e-16 ***
claimant_age                  -1.142e-02  9.756e-04 -11.705  < 2e-16 ***
GenderMale                     5.467e-02  2.311e-02   2.365  0.01802 *
GenderNot Available           -5.036e-01  1.632e-01  -3.086  0.00203 **
IsFatality                     5.326e-01  1.906e-01   2.795  0.00520 **
ClaimOpenDateOnWeek            1.533e-02  8.148e-03   1.882  0.05985 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54971  on 42125  degrees of freedom
Residual deviance: 45480  on 42110  degrees of freedom
AIC: 45512

Number of Fisher Scoring iterations: 7
```

```
> exp(coef(finalmodel))
                   (Intercept)                TotalIncurredCost          ClaimantTypeMedical Only
                     0.9818129                     1.0000883                     0.3851711
         ClaimantTypeReport Only   BodyPartRegionLower Extremities BodyPartRegionMultiple Body Parts
                     0.7489827                     1.0893462                     2.2160148
             BodyPartRegionNeck   BodyPartRegionNon-Standard Code              BodyPartRegionTrunk
                     0.9230626                     0.1436451                     1.3484362
  BodyPartRegionUpper Extremities                 recovery_period                    claimant_age
                     1.1429743                     1.0024368                     0.9886457
                    GenderMale             GenderNot Available                    IsFatality
                     1.0561923                     0.6043259                     1.7033092
            ClaimOpenDateOnWeek
                     1.0154526
```

Based on our final model for processing time, we believe the following attributes are statistically significant in predicting the probability of processing make the processing time critical. Total Incurred Cost, Claimant Type, Body Part Region, Recovery Period, Claimant Age, Gender, and Fatality.

- Claimant Type - Medical Only: Compared to the indemnity claimant type, the odds ratio of the medical only claimant type leading to a critical processing time is smaller by 0.38517 times. The coefficient of report only claimant type is not significant.
- Body Part Region: The coefficient is not significant when the injury body region is neck. For the injury body region is lower extremities, the odds ratio of processing time being critical is 1.089 times larger than the injury region is head. For the injury body region is multiple body parts, the odds ratio of processing time being critical is 2.216 times larger than the injury region is head. For the injury region is trunk, the odds ratio of processing time being critical is 1.348 times larger than the injury region is head. For the injury region is upper extremities, the odds ratio of processing time being critical is 1.1429 times larger than the injury region is head. Even though the coefficient of non-standard code is significant, it is not helpful for further analysis.
- Recovery Period: For every one day increase in the recovery day, the odds ratio of processing time being critical is increased by 1.0024 times.
- Claimant Age: For every one unit increase in claimant age, the odds ratio of processing time being critical is decreased by 0.9886 times.
- Gender: The odds of processing time being critical is 1.056 times larger for males than for females. For the unavailable gender, the odds ratio of processing time being critical is 0.604 times smaller than a female who has a critical processing time, however, this variable is not that useful for the analysis.
- Fatality: If the claim is fatality, the odds ratio of processing time being critical is increased by 1.7033 times.
- Even though the coefficient of Total Incurred Cost is significant to the criticalness of processing time, it does not lead to an obvious change in the probability of being critical.

According to the interpretation of model results, the most important factors are the body part region, the claim's fatality and claimant type. First, we need to set the head injury as a control group. Based on the control group, if a worker has multiple body parts injured and needs to file a claim, the probability of processing time being critical is increased by 69% which is the most important factor impacting the processing time. The next one is when the injured body part is the trunk. The probability of processing time being critical is increased by 57%. If the injured part is upper extremities, then the probability of processing time is increased by 53%. Lower extremities lead to 52.13% increase in processing time being critical. Another important factor is the fatality. If the claim is fatal, then the probability of processing time being critical is 63% larger than the non-fatal claim. The third factor we want to discuss is the claimant type. First, Indemnity type is set as the control group. If it is medical only, the probability of processing time being critical is decreased by 27.8%.

2. Total Incurred Cost as outcome variable

```
> summary(finalmodel)

Call:
glm(formula = PaymentBinary ~ ClaimantType + IndemnityPaid +
    TotalRecovery + day_difference + claimant_age + recovery_period +
    BodyPartRegion + ClaimOpenDateOnWeek, family = binomial,
    data = claimData)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.982  -0.942  -0.001   1.349   4.447

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.294e+00  6.904e-02  18.747  < 2e-16 ***
ClaimantTypeMedical Only       -2.048e+00  3.996e-02 -51.240  < 2e-16 ***
ClaimantTypeReport Only        -1.398e+01  6.224e+01  -0.225 0.822252
IndemnityPaid                   1.343e-03  9.108e-05  14.743  < 2e-16 ***
TotalRecovery                  -3.772e-04  3.646e-05 -10.345  < 2e-16 ***
day_difference                 -1.092e-04  1.502e-05  -7.270 3.60e-13 ***
claimant_age                    5.104e-03  9.469e-04   5.391 7.01e-08 ***
recovery_period                 6.291e-04  1.412e-04   4.455 8.40e-06 ***
BodyPartRegionLower Extremities 1.014e-01  4.048e-02   2.504 0.012272 *
BodyPartRegionMultiple Body Parts 2.314e-01 4.850e-02   4.771 1.84e-06 ***
BodyPartRegionNeck              1.547e-01  6.206e-02   2.492 0.012689 *
BodyPartRegionNon-Standard Code -1.096e-01 1.676e-01  -0.654 0.513125
BodyPartRegionTrunk             9.166e-02  4.346e-02   2.109 0.034936 *
BodyPartRegionUpper Extremities 5.218e-02  3.850e-02   1.355 0.175292
ClaimOpenDateOnWeek            -3.018e-02  8.017e-03  -3.764 0.000167 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58399  on 42125  degrees of freedom
Residual deviance: 45622  on 42111  degrees of freedom
AIC: 45652

Number of Fisher Scoring iterations: 11

> exp(coef(finalmodel))
                     (Intercept)          ClaimantTypeMedical Only
                    3.648409e+00                      1.290298e-01
                    IndemnityPaid                     TotalRecovery
                    1.001344e+00                      9.996229e-01
                     claimant_age                   recovery_period
                    1.005117e+00                      1.000629e+00
BodyPartRegionMultiple Body Parts                BodyPartRegionNeck
                    1.260309e+00                      1.167266e+00
             BodyPartRegionTrunk   BodyPartRegionUpper Extremities
                    1.095989e+00                      1.053565e+00
          ClaimantTypeReport Only
                    8.460798e-07
                   day_difference
                    9.998908e-01
BodyPartRegionLower Extremities
                    1.106691e+00
BodyPartRegionNon-Standard Code
                    8.962006e-01
              ClaimOpenDateOnWeek
                    9.702734e-01
```

For the final model of Payment, the following attributes are significant. Claimant Type, Indemnity Paid, Total Recovery, Processing Time, Claimant Age, Recovery Period, Body Part Region and Claim Open Date On Week.

- Claimant Type - Medical Only: Compared to the indemnity claimant type, the odds ratio of the medical only claimant type leading to a critical processing time is smaller by 0.1290 times. The coefficient of report only claimant type is not significant.
- Indemnity Paid: For every unit increase in indemnity paid, the odds ratio of the payment being critical is increased by 1.0013 times or by 0.13%, which will not lead to an obvious change in the probability of payment being critical.
- Total Recovery: For every one unit increase in the total recovery paid, the odds ratio of the payment being critical is decreased 0.9996 times. This attribute won't lead to an obvious change in the probability of payment being critical.
- Processing Time: For every one day increase in the processing time, the odds ratio of the payment being critical is decreased 0.9998 times. Also this attribute won't lead to an obvious change in the probability of payment being critical.
- Claimant Age: For every one unit increase in the claimant age, the odds ratio of the payment being critical is increased by 1.005. Again, this attribute won't lead to an obvious change in the probability of payment being critical.
- Recovery Period: For every one day increase in recovery period, the odds ratio of the payment being critical is increased by 1.0006 times. This attribute won't lead to an obvious change in the probability of payment being critical.
- Body Part Region: If the injury body part is neck, the odds ratio of payment being critical is 1.1672 times larger than the injury region is head. If the injury body region is lower extremities, the odds ratio of payment being critical is 1.1067 times larger than the injury region is head. For the injury body region is multiple body parts, the odds ratio of payment being critical is 1.2603 times larger than the injury region is head. For the injury region is trunk, the odds ratio of payment being critical is 1.0960 times larger than the injury region is head. The coefficient of non-standard code and upper extremities are not significant.
- Claim Open Date On Week: For every one day increase, the odds ratio of payment being critical is decreased by 0.9703 times.

According to the interpretation of model results, the most important factors impacting the payment are claimant type, injured body part region, and Claim Open Date on Week Day. The most important factor impacting the payment is Body Part Region. Again, we set Head as the control group. If the injured part is multiple body parts, the probability of payment being critical increased by 55.76%. If the injured part is the neck, the probability of payment being critical is increased by 53.86%. Lower extremities lead to 52.53% increase in probability. The increased percentage of the injured part is the trunk is 52.29%. For Claimant Type, first, we set Indemnity type as the control group. Then if the claimant type is medical only, the probability of payment being critical is decreased by 11.4%. The third important factor is the Claim Open Day on Week. The week range is from Monday to Friday. If the claim is opened one day later in a week, the probability of payment being critical is decreased by 49.25%.

Overall, based on the two models run for claim dataset, the most important factors impacting both payment and processing time are Body Part Region and Claimant Type. Multiple body parts were injured leads to increased probability in both payment and processing time being critical more than half. In contrast, medical only claimant type leads to decrease the probability. One special factor for processing time is the fatality. It leads to 63% increase in the probability of processing time being critical than non-fatality. Payment model also has its own specific factor which is Claim Open Day on Week. Most claims with critical payment have a higher probability to be filed on Monday.

Recommendations

In phase 1, we learned that total processing time was a major cost driver for the organization and its business activities. A mistake that we made in the first phase was that for all of our insights, we were only looking at the aggregated total incurred costs. Structuring our visualizations that way did not give us any insights into the data on a claims level basis. Our big picture goal for phase 2 is to help the company identify and eliminate inefficiencies in its business processes. In order to support this goal, we developed three strategic objectives that the company can use to measure the success or failure of an analytics project:

1) Minimize the money spent on storing and managing open claims by ensuring only legitimate open claims stay open
2) Proactively and efficiently allocate resources
3) Enhance our predictive capabilities by expanding the types of data being collected.

On an individual claim level, these objectives translate into a need to address a few key areas. We provide recommendations concerning how to treat claims involving multiple body parts and claims filed towards either end of the week, but it is important to first understand the relationship between reserve funds and administrative costs.

The administrative expense that we decided to focus on can be thought of as a technological overhead expense. The term technological overhead expense refers to the costs associated with storing and maintaining an open claim in our system. The money it takes for someone to examine an open claim in weekly report after weekly report is precious capital that could be used for something else. Another way of saying this is that when a claim is filed, addressed, and closed in a timely manner, it will require fewer administrative costs and improve the company's operations. It might seem that resolving these inefficiencies only slightly impacts our decision making and organizational effectiveness, but when you consider that the impact of those variables will improve decision making in more than 2 million claim instances, it is easy to see the economic value of such changes.

Reserve funds are used to pay expenses associated with a claim as they become due. So a having adequate reserve funds allocated to a claim before the expenses become due would mean that the claim could be closed as soon as it becomes possible, without having to wait for funds to be allocated. This means that if we are able to accurately predict where reserve

funds should be allocated, then we will experience a net decrease in administrative costs. This is important to note because even in our model predicting total incurred costs, we are only interested in the prediction so that we can allocate reserve funds preemptively- we are not directly reducing total incurred costs, only the administrative costs it takes to maintain an open claim in our system. Now that the underlying logic has been explained, we can take a look at the actual recommendations we have created for the company.

Claims involving multiple body parts performed as would have been expected- more likely to be associated with critical outcomes in processing time and total incurred costs. According to our first model where we set the outcome variable to processing time, the odds of a claim processing time being deemed critical increases more than 220% or is 2.2 times as likely in cases involving multiple body parts when compared to head injuries.

In our initial hypotheses, we expected claims filed on Mondays to act differently than the other days of the week. We figured that since Monday is the first day after a weekend, it would include claims from people who noticed their injuries on Friday evening after work, the next morning when they woke up, or even Sunday when the pain persisted. As we can see from the output of the payment logistic regression model, these initial thoughts were confirmed. The odds of the total cost incurred by a claim classifying as critical decreases as the week goes by. This is to say that claims filed towards the beginning of the week are probably going to be more expensive and therefore should be allocated more funds to cover costs as they become due. We recommend addressing this issue by raising the reserve requirements of claims that were filed on a Monday and for claims involving multiple body parts.

Contrarily, we found that the time to process a claim was more likely to be deemed critical if the claim was filed at the end of the week. So we recommend that the company expedite claims that have been predicted to score critically for processing time. In other words, claims that were filed on Friday should be brought to the front of the hypothetical “to-do” list so that these at-risk claims can be addressed promptly. Doing so would ensure that we are only keeping claims open for legitimate reasons rather than letting them linger and eat up administrative costs. However, not all instances of processing time can be reduced by improving a business’ process.

For example, we found that claims related to a workplace fatality are more likely to score critically in relation to processing time. It stands to reason that deaths tend to indicate a higher time to process because the process is held up by constraints that are not present in the case of a broken leg. The need to wait for things like official death certificates, liability rulings, etc. to verify the veracity of the claims involving workplace deaths make it a tough issue to address. Even so, the attribute was found to be significant, both statistically and in regards to its impact on our predictions, and it should therefore be included in future models.

In addition to the data that was included in our initial and wrangled datasets, we recommend the client company start collecting data for expected recovery time and to derive more variables to achieve even greater predictive capability of our models. High recovery

periods were associated with both critical time processing claims and total costs incurred by a claim. However, a longer recovery period did not necessarily ‘cause’ an increase in the odds of a critical outcome which reminded us of a phrase common in discussing statistical output- “correlation does not mean causation.” We found that a 1 unit increase in recovery time leads to a nearly identical change in the probability of a critical outcome. So if we are able to get expected recovery times that are even partially accurate, we can improve the predictive capability of our models in the future. A derived variable that would indicate whether an employee returned to work before their claim was closed would also improve our ability to predict key claims that should be expedited in our system. Moving forward, the claims processing company should implement our recommendations, continue to monitor the results to tweak the model periodically, and work to expand the types of data they are collecting and analyzing so that other modeling techniques can be used in the future which could provide us with insights that logistic regression cannot provide us with.

Appendix 1 – New Variables Created

1. **Claim Open Day On Week**: we obtain this column by switching each date to the day of the week, aiming to see if particular working day has more significant on processing time/ payment.
2. **Date difference/Processing Time**: derived from “Claimant open date” minus “Claimant closed date”
3. **Recovery period** : shows how long it between” Incident date” and “Return to work date”, and we use that to test if it matters the payment amount or processing time.

Appendix 2 – Model Output

1. Processing Time

```
> summary(finalmodel)

Call:
glm(formula = TimeBinary ~ TotalIncurredCost + ClaimantType +
    BodyPartRegion + recovery_period + claimant_age + Gender +
    IsFatality + ClaimOpenDateOnWeek, family = binomial, data = claimData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8288  -0.7685  -0.6838   0.9438   2.6457

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.835e-02  6.713e-02  -0.273  0.78455
TotalIncurredCost               8.832e-05  3.233e-06  27.319  < 2e-16 ***
ClaimantTypeMedical Only       -9.541e-01  3.082e-02 -30.956  < 2e-16 ***
ClaimantTypeReport Only        -2.890e-01  4.016e-01  -0.720  0.47168
BodyPartRegionLower Extremities 8.558e-02  4.339e-02   1.972  0.04856 *
BodyPartRegionMultiple Body Parts 7.957e-01  4.910e-02  16.206  < 2e-16 ***
BodyPartRegionNeck             -8.006e-02  6.739e-02  -1.188  0.23486
BodyPartRegionNon-Standard Code -1.940e+00  3.357e-01  -5.781 7.43e-09 ***
BodyPartRegionTrunk             2.989e-01  4.539e-02   6.587 4.49e-11 ***
BodyPartRegionUpper Extremities 1.336e-01  4.140e-02   3.228  0.00125 **
recovery_period                 2.434e-03  2.131e-04  11.419  < 2e-16 ***
claimant_age                   -1.142e-02  9.756e-04 -11.705  < 2e-16 ***
GenderMale                      5.467e-02  2.311e-02   2.365  0.01802 *
GenderNot Available            -5.036e-01  1.632e-01  -3.086  0.00203 **
IsFatality                      5.326e-01  1.906e-01   2.795  0.00520 **
ClaimOpenDateOnWeek             1.533e-02  8.148e-03   1.882  0.05985 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54971  on 42125  degrees of freedom
Residual deviance: 45480  on 42110  degrees of freedom
AIC: 45512

Number of Fisher Scoring iterations: 7
```

```
> exp(coef(finalmodel))
                      (Intercept)                 TotalIncurredCost          ClaimantTypeMedical Only
                        0.9818129                         1.0000883                         0.3851711
          ClaimantTypeReport Only   BodyPartRegionLower Extremities BodyPartRegionMultiple Body Parts
                        0.7489827                         1.0893462                         2.2160148
               BodyPartRegionNeck   BodyPartRegionNon-Standard Code               BodyPartRegionTrunk
                        0.9230626                         0.1436451                         1.3484362
  BodyPartRegionUpper Extremities                   recovery_period                      claimant_age
                        1.1429743                         1.0024368                         0.9886457
                       GenderMale              GenderNot Available                       IsFatality
                        1.0561923                         0.6043259                         1.7033092
             ClaimOpenDateOnWeek
                        1.0154526

> n <- nrow(claimData)
> logLik(finalmodel)
'log Lik.' -22740.02 (df=16)
> lrtest(logitbase, finalmodel)
Likelihood ratio test

Model 1: TimeBinary ~ 1
Model 2: TimeBinary ~ TotalIncurredCost + ClaimantType + BodyPartRegion +
    recovery_period + claimant_age + Gender + IsFatality + ClaimOpenDateOnWeek
  #Df LogLik Df  Chisq Pr(>Chisq)
1   1 -27486
2  16 -22740 15 9491.4  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> McfadR2 <- 1-((finalmodel$deviance/-2)/(finalmodel$null.deviance/-2))
> cat("mcFadden R2=",McfadR2,"\n")
mcFadden R2= 0.172661
> AIC<- finalmodel$deviance+2*2
> cat("AIC=",AIC,"\n")
AIC= 45484.05
> pR2(finalmodel)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-2.274002e+04 -2.748574e+04  9.491432e+03  1.726610e-01  2.017317e-01  2.767966e-01
> predprob1 <- fitted(finalmodel)
> probTable1 <- data.frame(predprob1)
> table(predprob1>.5, claimData$TimeBinary)

            0     1
  FALSE 25132  8875
  TRUE   1898  6221
```

2. Payment

```
> summary(finalmodel)

Call:
glm(formula = PaymentBinary ~ ClaimantType + IndemnityPaid +
    TotalRecovery + day_difference + claimant_age + recovery_period +
    BodyPartRegion + ClaimOpenDateOnWeek, family = binomial,
    data = claimData)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.982  -0.942  -0.001   1.349   4.447

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.294e+00  6.904e-02  18.747  < 2e-16 ***
ClaimantTypeMedical Only        -2.048e+00  3.996e-02 -51.240  < 2e-16 ***
ClaimantTypeReport Only         -1.398e+01  6.224e+01  -0.225 0.822252
IndemnityPaid                    1.343e-03  9.108e-05  14.743  < 2e-16 ***
TotalRecovery                   -3.772e-04  3.646e-05 -10.345  < 2e-16 ***
day_difference                  -1.092e-04  1.502e-05  -7.270 3.60e-13 ***
claimant_age                     5.104e-03  9.469e-04   5.391 7.01e-08 ***
recovery_period                  6.291e-04  1.412e-04   4.455 8.40e-06 ***
BodyPartRegionLower Extremities  1.014e-01  4.048e-02   2.504 0.012272 *
BodyPartRegionMultiple Body Parts 2.314e-01  4.850e-02   4.771 1.84e-06 ***
BodyPartRegionNeck               1.547e-01  6.206e-02   2.492 0.012689 *
BodyPartRegionNon-Standard Code -1.096e-01  1.676e-01  -0.654 0.513125
BodyPartRegionTrunk              9.166e-02  4.346e-02   2.109 0.034936 *
BodyPartRegionUpper Extremities  5.218e-02  3.850e-02   1.355 0.175292
ClaimOpenDateOnWeek             -3.018e-02  8.017e-03  -3.764 0.000167 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58399  on 42125  degrees of freedom
Residual deviance: 45622  on 42111  degrees of freedom
AIC: 45652

Number of Fisher Scoring iterations: 11

> exp(coef(finalmodel))
                     (Intercept)          ClaimantTypeMedical Only
                    3.648409e+00                      1.290298e-01
                    IndemnityPaid                     TotalRecovery
                    1.001344e+00                      9.996229e-01
                     claimant_age                   recovery_period
                    1.005117e+00                      1.000629e+00
BodyPartRegionMultiple Body Parts                BodyPartRegionNeck
                    1.260309e+00                      1.167266e+00
              BodyPartRegionTrunk   BodyPartRegionUpper Extremities
                    1.095989e+00                      1.053565e+00
          ClaimantTypeReport Only                     day_difference
                    8.460798e-07                      9.998908e-01
  BodyPartRegionLower Extremities                  BodyPartRegionNon-Standard Code
                    1.106691e+00                      8.962006e-01
              ClaimOpenDateOnWeek
                    9.702734e-01
```

```
> n <- nrow(claimData)
> logLik(finalmodel)
'log Lik.' -22811.17 (df=15)
> lrtest(logitbase, finalmodel)
Likelihood ratio test

Model 1: PaymentBinary ~ 1
Model 2: PaymentBinary ~ ClaimantType + IndemnityPaid + TotalRecovery +
    day_difference + claimant_age + recovery_period + BodyPartRegion +
    ClaimOpenDateOnWeek
  #Df LogLik Df Chisq Pr(>Chisq)
1   1 -29200
2  15 -22811 14 12777  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> McfadR2 <- 1-((finalmodel$deviance/-2)/(finalmodel$null.deviance/-2))
> cat("mcFadden R2=",McfadR2,"\n")
mcFadden R2= 0.2187827
> AIC<- finalmodel$deviance+2*2
> cat("AIC=",AIC,"\n")
AIC= 45626.34
> pR2(finalmodel)
         llh       llhNull            G2      McFadden          r2ML          r2CU
-2.281117e+04 -2.919952e+04  1.277670e+04  2.187827e-01  2.616204e-01  3.488272e-01
> predprob1 <- fitted(finalmodel)
> probTable1 <- data.frame(predprob1)
> table(predprob1>.5, claimData$PaymentBinary)

           0     1
  FALSE 20062 11074
  TRUE   1001  9989
```

Appendix 3 – Binary Dependent Variables' Basis

1. Based on total payments
    a. Median of total incurred costs (which is $437.775). We set the value higher than the median as 1 (critical), the value lower or equal than the median as 0 (noncritical).
2. Based on processing time
    a. 12 months as the divider based on our research. Processing time longer than 12 months as 1 (critical), processing time shorter or equal than 12 months as 0 (noncritical). Attached the screenshot.

Appendix 4 – SAS Result for Double Checking

1. Processing Time

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	ClaimantType	2	1	6361.6270	<.0001	
2	TotalIncurredCost	1	2	709.6104	<.0001	
3	BodyPartRegion	6	3	503.5156	<.0001	
4	recovery_period	1	4	175.6643	<.0001	
5	claimant_age	1	5	138.9541	<.0001	
6	Gender	2	6	17.1532	0.0002	
7	IsFatality	1	7	7.8452	0.0051	

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
TotalIncurredCost	1	745.7753	<.0001
claimant_age	1	136.8243	<.0001
recovery_period	1	130.4737	<.0001
Gender	2	16.4327	0.0003
ClaimantType	2	960.6195	<.0001
IsFatality	1	7.7606	0.0053
BodyPartRegion	6	467.3768	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1018	0.4743	0.0461	0.8300
TotalIncurredCost		1	0.000088	3.231E-6	745.7753	<.0001
claimant_age		1	-0.0114	0.000976	136.8243	<.0001
recovery_period		1	0.00243	0.000213	130.4737	<.0001
Gender	Female	1	0.5058	0.1632	9.6094	0.0019
Gender	Male	1	0.5610	0.1632	11.8189	0.0006
ClaimantType	Indemnity	1	0.2885	0.4015	0.5163	0.4724
ClaimantType	Medical Only	1	-0.6658	0.4007	2.7603	0.0966
IsFatality	0	1	-0.5309	0.1906	7.7606	0.0053
BodyPartRegion	Head	1	-0.1331	0.0414	10.3311	0.0013
BodyPartRegion	Lower Extremities	1	-0.0481	0.0311	2.3935	0.1218
BodyPartRegion	Multiple Body Parts	1	0.6623	0.0387	292.1175	<.0001
BodyPartRegion	Neck	1	-0.2141	0.0603	12.6053	0.0004
BodyPartRegion	Non-Standard Code	1	-2.0752	0.3344	38.5009	<.0001
BodyPartRegion	Trunk	1	0.1652	0.0340	23.6671	<.0001

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
TotalIncurredCost	1.0000	1.000	1.000	1.000
claimant_age	1.0000	0.989	0.987	0.991
recovery_period	1.0000	1.002	1.002	1.003
Gender Female vs Not Available	1.0000	1.658	1.204	2.283
Gender Male vs Not Available	1.0000	1.752	1.273	2.413
ClaimantType Indemnity vs Report Only	1.0000	1.334	0.607	2.931
ClaimantType Medical Only vs Report Only	1.0000	0.514	0.234	1.127
IsFatality 0 vs 1	1.0000	0.588	0.405	0.854
BodyPartRegion Head vs Upper Extremities	1.0000	0.875	0.807	0.949
BodyPartRegion Lower Extremities vs Upper Extremities	1.0000	0.953	0.897	1.013
BodyPartRegion Multiple Body Parts vs Upper Extremities	1.0000	1.939	1.797	2.092
BodyPartRegion Neck vs Upper Extremities	1.0000	0.807	0.717	0.909
BodyPartRegion Non-Standard Code vs Upper Extremities	1.0000	0.126	0.065	0.242
BodyPartRegion Trunk vs Upper Extremities	1.0000	1.180	1.104	1.261

2. Payment

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ClaimOpenDateOnWeek	1	15.7666	<.0001
ClaimantType	2	5423.3003	<.0001
BodyPartRegion	6	24.3412	0.0005

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-12.3239	85.7150	0.0207	0.8857
day_difference		1	-0.00008	0.000015	32.2766	<.0001
recovery_period		1	0.00170	0.000176	93.6932	<.0001
claimant_age		1	0.00646	0.000944	46.8590	<.0001
ClaimOpenDateOnWeek		1	-0.0316	0.00797	15.7666	<.0001
ClaimantType	Indemnity	1	14.3422	85.7150	0.0280	0.8671
ClaimantType	Medical Only	1	11.5729	85.7150	0.0182	0.8926
BodyPartRegion	Head	1	-0.0992	0.0386	6.5956	0.0102
BodyPartRegion	Lower Extremities	1	0.0449	0.0299	2.2568	0.1330
BodyPartRegion	Multiple Body Parts	1	0.1167	0.0402	8.4466	0.0037
BodyPartRegion	Neck	1	0.0253	0.0564	0.2008	0.6540
BodyPartRegion	Non-Standard Code	1	-0.2007	0.1668	1.4467	0.2291
BodyPartRegion	Trunk	1	-0.00555	0.0340	0.0267	0.8703

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	73.5	Somers' D	0.470
Percent Discordant	26.5	Gamma	0.470
Percent Tied	0.0	Tau-a	0.235
Pairs	443649969	c	0.735

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
day_difference	1.0000	1.000	1.000	1.000
recovery_period	1.0000	1.002	1.001	1.002
claimant_age	1.0000	1.006	1.005	1.008
ClaimOpenDateOnWeek	1.0000	0.969	0.954	0.984
ClaimantType Indemnity vs Report Only	1.0000	>999.999	<0.001	>999.999
ClaimantType Medical Only vs Report Only	1.0000	>999.999	<0.001	>999.999
BodyPartRegion Head vs Upper Extremities	1.0000	0.906	0.839	0.977
BodyPartRegion Lower Extremities vs Upper Extremities	1.0000	1.046	0.986	1.109
BodyPartRegion Multiple Body Parts vs Upper Extremities	1.0000	1.124	1.039	1.216
BodyPartRegion Neck vs Upper Extremities	1.0000	1.026	0.918	1.146
BodyPartRegion Non-Standard Code vs Upper Extremities	1.0000	0.818	0.590	1.135