University of North Carolina at Charlotte

**Group Project Phase #2**

Jiawei Ma

Xing Cai

Omkar Bapat

Vishak Lakshman SM

Sai Shanoor

DSBA 6100

Professor Chandra Subramaniam

12/06/2017

**EXECUTIVE SUMMARY**

**OBJECTIVE:**

The main objective is to find the major drivers of the costs and processing time of the claims, and find the ways to reduce or avoid the impact of the factors, and make particular recommendations for our client. To reach this goal we need to bring our data and ideas from the phase 1, and to use the model to acquire more detailed information and useful facts. Finally we can use these information to provide more useful strategies to improve the competitive advantages for our client company.

**INITIAL HYPOTHESES:**

* We predicted that the older claimant will be more likely having the higher cost for a claim.
* If a claim type is indemnity, the payment and processing time will be higher than the cases are medical only or report only.
* Claims with higher total days of duration, tend to have higher amount of total paid.
* The payment and processing time are higher, when the cases are denied or fatality.

**REPORT**

**MODELS MAY BE USED:**

For the multiple linear regression model, we can use it to predict the cost of a compensation claim or the processing time of a claim with the given the key factors. We also can be better understand the relationship between dependent variable and independent variables. We also can find that which variable is the biggest driver of the cost and processing time for a claim, and what the magnitude of the impact. But the linear regression also has the limitations. For the linear regression, the dependent variable and independent variables have to be the linear relationships. And all the independent variables have to be independent for each other. And the variance of residuals have to be equal, so the linear regression may not work for a lot of data. We construct two multilinear regression models. One is to predict the processing time and the second one is to predict cost:

* + Dependent variables: Claim\_Duration
  + Independent variables:
    - SUM\_TrancPayment (we believe more amount of payment paid, the longer duration it has)
    - ClaimantAge\_at\_DOI (we want to see what is the impact of age on processing time)
    - IsDenied (we want to see how much longer the processing time will be if the claim is being denied)
    - Gender (we want to compare the difference of claim duration between male and female)
    - ClaimantType (to understand which type of claim has the longer duration may be helpful for the decisions)
    - BodyPartRegion (we want to compare with the time duration for different body parts)
    - Smallinjury (we want to know how big the difference is between small common injuries such as strain, sprain and the other injuries)
  + Dependent variables: SUM\_TrancPayment
  + Independent variables:
    - ClaimantAge\_at\_DOI (we found the claimant age and cost have positive relationship during the visualization, so we want to prove it)
    - IsDenied (we want to compare which one has higher cost, denied claims or not denied)
    - Gender (we want to compare cost between male and female)
    - ClaimantType (to understand which type of claim has the higher cost may be helpful for the decisions)
    - BodyPartRegion (we want to know which part of the body gets higher cost)
    - Claim\_Duration (we want to see the relationship between cost and claim duration)
    - Smallinjury (because the small injuries are the most frequent injuries, we want to know the gap of cost between small injuries and the other injuries)

Logistic Regression model is searching for a single linear decision boundary in the feature space. Logistic regression model enables us to estimate the probability of a high cost claim will occur or the probability of a long duration claim will occur. To better perform the logistic regression model, we create two new binary variables, HighPayment and LongDuration. HighPayment indicates the SUM\_TrancPayment>$700 and LongDuration is where the Claim\_Duration greater than 180 days. And logistic regression can also help us to better understand which factors will increase or decrease the probability of high cost claim or long duration claim occur, by how much it will increase or decrease. However the limitation for logistic regression is that the dependent variable has to be the binary variable, otherwise it will not work. Here are the two logistic regressions we construct:

* + Dependent variables: LongDuration
  + Independent variables:
    - SUM\_TrancPayment (we want to know how much the probabilities of a long duration claim will increase or decrease, when cost increase)
    - ClaimantAge\_at\_DOI (we want to see whether age affects the probabilities of a long duration claim occurring)
    - IsDenied (we think the denied claims are more likely to be longer processing time)
    - Gender (we want to compare male and female in this case)
    - ClaimantType (we want to know which clam type has higher chance to be the long duration claim)
    - BodyPartRegion (we think different body parts have different impact)
    - Smallinjury (we want to compare the small injuries and the other injuries)
  + Dependent variables: HighPayment
  + Independent variables:
    - ClaimantAge\_at\_DOI (we want to know how much the probabilities of a expensive claim will increase or decrease, when claimant age increases)
    - IsDenied (we think the denied claims are more likely to be more expensive claim)
    - Gender (we want to which gender is more likely to become more expensive claim)
    - ClaimantType (to know which claim type is more likely to be the expensive claim may be helpful)
    - BodyPartRegion (to know which body part involves a claim could affect the probabilities of being expensive claim may be helpful)
    - Smallinjury (we want to know whether small injuries can have higher chance to be the expensive claim)
    - Claim\_Duration (we want to know how much the probabilities of an expensive claim occur will increase or decrease, when duration increase)

A decision tree is essentially partitioning the feature space into half-spaces using *axis-aligned* linear decision boundaries. It is a classification technique to group observations into different categories. We can use the decision tree to divide the claims into different categories and find the possible consequences. Similar with the logistic regression, decision tree also allows us to find the chance of outcomes. Decision tree also has limitation. For example, if a variable has too many branches or choices, such as BodyPartRegion, then it will not work well.

* Decision Trees:
  + Dependent variables: HighPayment
  + Independent variables: IsDenied, IsFatality, Gender, ClaimantType

Those variables are easy to divide into different categories, and few branches, more accuracy.

In this situation, there are about thirty variables can be used to analyze, including categorical variables with lots of subcategories and some continuous variables. Applying Decision Tree model to this problem may cause overfitting and get lots of nods and branches which are difficult to explain. With Logistic Regression model, we can easily get the most important drivers of claims payments and processing time.

**Answers to the hypotheses:**

* The first hypothesis we made is that the older claimant will be more likely having the higher cost for a claim.