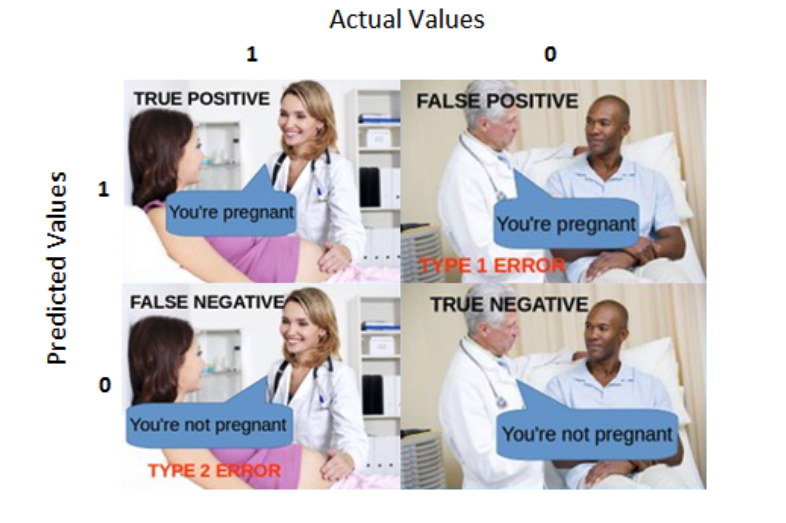
TOC

1. Introduction
2. Data and Cleaning Process
3. Supervised Machine Learning Methods Applied
4. Analysis and Model Selection
5. Results of Analysis
6. Conclusion
7. Appendix

**Introduction**

* Summarize the purpose of the report and summarize the data / subject.
  + This report will provide our methods, model, analysis, and conclusions for the unlabeled data set provided. Along with that, the data cleaning process will be explained and details on our findings provided.
  + To recap, we were assigned a dataset with 160,000 records and 51 features. According to our business partner, the data is related to the insurance industry. For the most part, the features were unlabeled – meaning the data has not been tagged identifying its characteristics, properties, or classifications. However, the target variable (i.e. dependent variable) is labeled (0 or 1). As a result, we will be looking at various supervised machine learning methods to arrive at our conclusions.
  + For general background, supervised machine learning (SML) learns from the dataset to for classification and/or regression purposes. In this report, we will use the following methods for classification:
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Support Vector Machine
    - K- Nearest Neighbor
  + The objective given by our business partner is to arrive at a model that best classifies the ‘y’ values (0 or 1). Hence, this is a classification problem and our conclusions will report the model with the best accuracy in classifying records/observations.
  + What does classification mean? At its core, classification learns from the data given (i.e. training dataset) and then uses the findings to create a model to accurately classify new records. Along with accuracy, we will show the ratio of actual values to predicted values. As advised by our business partner, it is important to the business to accurately classify but even more so to lower the amount of ‘False Positives’ – mis-classifying a false positive could cost the business $1,000 while mis-classifying a false negative $100. Below is a graphic to help explain the possible outcomes:



* Summary of Conclusion: Our recommendation is using K-NN with feature selection to best classify for this data set. Based on receiving new data with these same categories, the business will be able to predict/classify with approximately a 95.12% accuracy whether the transaction will result in money lost (0) or gained (1). Out of the 5% error rate in prediction, 4.07% were false positive results and 6.56% were false negative results.
* The remainder of the report will provide further details on our process of data cleaning, feature selection/engineering, methods, modeling, and analysis.

**Body - Four Sections**

* Data Section - Include written descriptions of data and follow with relevant spreadsheets.
  + Random Forest Feature Importance (RFFI)
    - Knowing feature importance allowed us to better understand the data set and provided insights to our feature selection process. After removing the categorical features, we ran RFFI for the remaining 45 features (using all 160,000 records) to arrive at the following feature of importance (>0.04):
      * ['x23', 'x20', 'x48', 'x49', 'x38', 'x12', 'x42', 'x27','x40', 'x37','x28','x7','x2', 'x46']
* Methods Section - Explain how you gathered and analyzed data.
  + We started to evaluate the SML methods stated above from the most common/basic to more complex. Each method selected utilized two versions of the data set to arrive at the method that yielded the best model (i.e. most accurate). The following will provide further detail into our process in sequential order of evaluation:
    - Logistic Regression (LR)
      * This method was the most obvious to run first since the dependent variable (‘y) was binary (0 or 1). The objective is to describe data and to explain the relationship between the dependent variable to the rest of the data set. Unfortunately, LR yielded the lowest accuracy rates in predicting the classification.
        + Full Data Set: 70.27%
        + Feature Selected: 70.33%
    - Decision Tree (DT)
      * The goal of DT is to split the observations in a way that the resulting groups are as different from each other as much as possible. The structure of a tree is the root node, internal node, and leaf nodes – all of which are connected by branches. DT is the building block for Random Forest. The DT yielded better accuracy results compared to LR.
        + Full Data: 84.09% accuracy using entropy
        + Feature Selected: 86.5% accuracy using entropy criterion.
    - Random Forest (RF)
      * RF uses a large number of uncorrelated decision trees to operate as a committee – much like an ensemble. Each DT outputs a class prediction and the class with the most votes is the RF’s optimal model prediction. The RF yielded better accuracy results compared to both LR and DT.
        + Full Data: 88.7% accuracy
        + Feature Selected: 91.5% accuracy
    - SVM
    - PCA
    - K-Nearest Neighbor (KNN)
      * KNN is a non-parametric, lazy learning machine learning algorithm with the purpose to predict the classification of data. Non-parametric means that no assumptions are made about the data before classifying; lazy learning means that there is no training phase before classifying. KNN often has the ‘dimensionality curse’, where a dataset with many variables is not as accurate; however, because our dataset is so large, we reduce the dimensionality curse. KNN is our ‘winning’ algorithm (see Analysis section for more details on the process). Other methods tried with KNN were adjusting the test size, one-hot-encoding all categorical variables, and using
        + Full Data: 80.24
        + Feature Selected:
* Analysis and Model Section - Explain what you analyzed. Include any charts here.
  + With KNN, we started with the full dataset and k=5. After yielding somewhat high results on the first pass at the full dataset, we ran a KNN loop (with just the feature selected data) from k=1=k=25. Results can be seen below:
    - INSERT ACCURACY KNN CHART HERE
  + The best accuracy was found at k=9, although not much variation occurred between k=5 and k=9. After k=9 the accuracy slowly dropped. Upon running KNN with the featured dataset, with a test size of 20% (32,000 samples), and a random\_state set at “1234”, we found the accuracy to be 95.17%.
    - INSERT CONFUSION MATRIX HERE
* Results - Describe the results of your analysis.
  + KNN had an accuracy of 95.17%. Within those results, a positive was correctly predicted 96.26% of the time, and a negative was correctly predicted 93.55% of the time. Incorrect classifications of positive results occurred 3.74% of the time, while incorrect classifications of negative results occurred 6.45% of the time.

**Conclusions**

* Restate the questions from your introduction.
* Restate important results.
  + You can expect to correctly predict a positive (y=1) outcome 96.17% of the time and correctly predict a negative (y=0) outcome 93.55% of the time.
* Include any recommendations for additional data as needed.
  + Further investigations can be done by analyzing the data with the categorical variables (continents, month, day). These datapoints did not prove significant in our findings, however, they may be important to the business partner.

**Appendix**

* Include the details of your data and process here.
* Include any secondary data, including references.