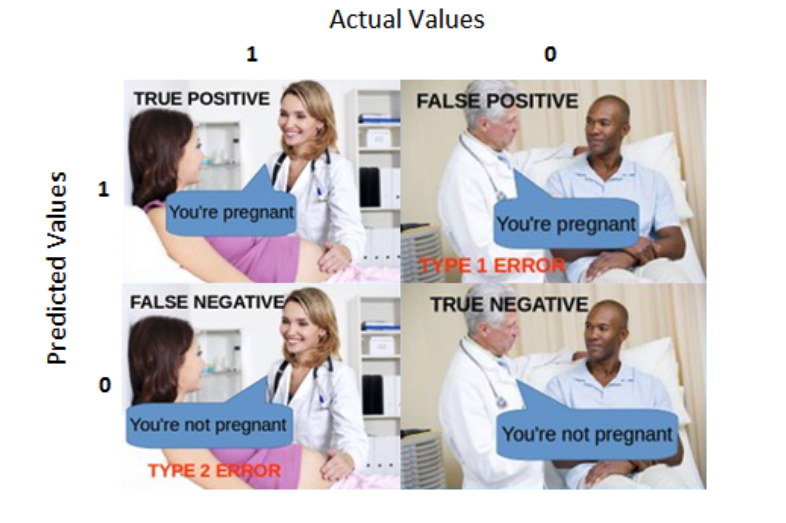
TOC

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 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 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% 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.
* Methods Section - Explain how you gathered and analyzed data.
* Analysis and Model Section - Explain what you analyzed. Include any charts here.
* Results - Describe the results of your analysis.

**Conclusions**

* Restate the questions from your introduction.
* Restate important results.
* Include any recommendations for additional data as needed.

**Appendix**

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