Final Case Study

MSDS Fall 2019

7333 Quantifying the World

Brandon de la Houssaye, Bruce Granger, Daniel Serna

# Abstract

In this paper, we demonstrate the use of multiple machine learning techniques to garner insight from an obfuscated data set. Despite being provided little context into the provided data, we are able to produce business value by leveraging a variety of traditional classifiers along with neural networks. Through comparison of different approaches, we recommend an algorithm that achieves effective accuracy metrics while minimizing false positives and false negatives.

# 1 Introduction

We have been approached with an interesting problem to solve. Our business stakeholders would like us to consume a dataset that has had domain context removed. All feature column names have been obfuscated, leaving it up to the team to garner insight as best we can. Ultimately, we have been tasked with building a model that can accurately predict the value of a column labeled “y.”

Through consultation with our business stakeholders, we have gained some minor context around the provided data set. At a high level, we know the data relates to insurance claims in some manner. We also know that our model predictions have monetary impact to the business. Most importantly, bad predictions cost the business money. Additionally, we have also been informed that false positives incur a 10x greater cost than false negatives. Thus, while achieving high accuracy with our model is desirable, we must ensure we minimize our false predictions and specifically focus on limiting false positives as much as possible. Because of this constraint, the recall metric of our models becomes important for analysis.

# 2 Data

As mentioned above, the data provided has been obfuscated by having all feature column names removed. The data contains fifty feature columns labeled x0 through x49 and one prediction column labeled y. Of the fitty feature columns, forty-seven are numeric variables and three are categorical variables. We are dealing with a binary prediction variable as column y only contains the values 1 or 0.

The dataset contains 160,000 records and for the most part is well formed. Fortunately, we do not have much missing data to deal with as can be seen in Figure 1 (note we have renamed some of the original columns).



Figure - Missing Data

As can be seen from Figure 1, all of our columns have less than fifty missing data points. When compared to the total record count of 160,000, this is almost a negligible number of missing values (<0.03%). We discuss our handling of the missing values in the Data Cleansing section.

Our data also appears to follow normal distributions and no significant deviations from normality were observed as can be from the boxplots in Figure 2.



Figure - Boxplots

As can be seen in the figure of above, we do not see evidence of any left or right skewness in the distributions. Two columns (x6 and x47) appear to have a wider distribution than the others, but still appear to follow a normal distribution. The boxplots also show little evidence of outliers. Thus, no transformations of the original data were deemed necessary.

We do have some collinearity in the data as can be seen from the Pearson correlation matrix in Figure 3.



Figure - Pearson Correlation

As can be seen in Figure 3, columns x2 and x6 are highly correlated with each other and columns x38 and x41 are highly correlated with each other. We will discuss our handling of these correlated columns in the Data Cleansing section.

# 3 Data Cleansing

# 4 Methods

# 5 Results

# 6 Conclusions