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.

# 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. 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.

# Data

# Data Cleansing

# Methods

# Results

# Conclusions