HW5 Proposal

RJ Cass

1. Understanding the Problem

Background:

We have a dataset of credit card transactions, including several redacted fields and the indication of whether or not the transaction is known to be fraudulent. We want to use this data to create a predictive model to identify which transactions are fraudulent or not for use by our legal team.

Goals:

In this analysis, we want to address several key questions:

i. How accurately can you identify the fraudulent transactions? In other words, given a transaction is fraudulent, how well do you identify it as such?

- To answer this we will create a predictive model using the given data. We will then use this model for 'In-Sample' prediction, and identify, out of the known fraudulent transactions, how many we are able to correctly identify.

ii. Which of the 5 sample transactions, if any, do you think are fraudulent?

- To answer this we will use the same predictive model created for the first question and apply it to the 'Out-of-Sample' data and predict whether they are fraudulent or not.

2. Exploratory Data Analysis

The first thing to consider in this dataset is that our output variable is binary (yes/no). As such, we will need to create our model based on the likelihood of the output. We also need to consider if all the continuous variables are monotonic (always increasing/decreasing) in relation to the likelihood of the outcome. Furthermore, there is an extremely heavy imbalance in the outcome result, where only .2% of the transactions are labelled as fraudulent, giving us a very limited amount of data to determine what indicates fraudulence. If these are not accounted for, our model will not be able to correctly predict, or it may perform well In-Sample, but poorly Out-of-Sample.

3. Desired Attributes

Model attributes required from the research questions

The research questions require that the model be able to predict accurately (in terms of which transactions are fraudulent) both in-sample and out-of-sample.

Model attributes required from the data

The data require that our model account for a binary outcome variable. They also require that we account for the heavy imbalance of outcome.

Any other anticipated problems

In our initial exploration, we identified that not all of the variables appear to be monotonic (always increasing or decreasing) in fraudulence. The model will either need to account for this, or we will need to use a model that does not require this.

What goes wrong if the above are not accounted for?

If all of our assumptions are not met, the resultant model will not correctly account for each factor, meaning the predictive power will decrease, or we may be able to predict very well In-Sample, but have poor Out-of-Sample performance.

4. Proposed Method

Appropriate models

Models appropriate for this analysis include a logistic regression, a KNN (K-nearest neighbors), and decision trees. In these cases, we can use Downsampling on the 'Not Fraudulent' transactions, and Upsampling on the 'Fraudulent' transactions.

Specific model proposal

For this analysis, we propose using a Random Forest. These models do not have a specific mathematical representation, but can be thought of as a binary tree where each branch represents a certain cutoff in the data, and can be used to identify which combination of factors indicates a transaction is fraudulent.

Method strengths/weaknesses

i. How this method accounts for issues in the dataset

This model does not have any model assumptions, meaning that the previously identified issues (such as non-monotinicity) are not relevant.

ii. How this method accomplishes research/analysis goals and yields appropriate estimators

This model predicts and will provide indicators for each data point as to whether or not it is fraudulent.

iii. How this method will answer the research questions

This model allows for prediction and will allow us to test In-Sample predictive power, and make predictions for Out-of-Sample data.

iv. What assumptions are needed to use the model adequately? Are they reasonable to assume and explained well?

This model does not have any inherent assumptions besides independence of each data point, which we are assuming is the case.