Marek Krawczyk

Milestone 1

DAT-690

9/3/2022

**Business Understanding Phase:** Before formulating an extensive and detailed plan, it is important to first grasp the problem we are facing. The Credit Project One: Loan default requires us to create a model that will be capable of accurately (MAPE < 30%) identifying customer data to predict if the credit they wish to extend is most likely to be defaulted on. This project faces many problems, from poor predictions, incorrect variable selection, and wrong model choices, but first and foremost we must focus on the overall goal of making meaningful, accurate, and sustainable predictions in business context of our enterprise.

**Research Question:**

Can an accurate (MAPE < 30%) model be created, based on consumer data voluntarily given by our prospective consumers, to predict the risk of their possible loan default?

**Business Solutions:**

Developing an advanced tool to be able to gauge the relative risk of loan defaults, bring s immense value to our enterprise. Having a support tool, that every loan officer can default to, in order to get data-based understanding of the risk profile, will reduce the amount of bad-debt by a significant amount. The training process and the experience threshold will be lowered, as the need for individual customer analysis, and expert level of experience of staff required. The customer side will see an improvement in loan approval turnaround times, and as the project matures, we will be able to ask only the relative questions, decreasing the input necessary from customer side. Another great opportunity lies in being able to tackle previously untapped markets from previously mislabeled or judged unsuitable demographics. Getting rid of demographic bias will help to improve our baseline and help communities in need.

**Data Understanding Phase:**

This next part is extremely important, understanding, verifying, cleaning and organizing data is majority of any data-science project. To structure and feed the data accordingly to what the model needs is the true workload.

Our Credit Project One: Loan default data has 31 variables, with a numerical index to count the number of observations. Total amount of observations is 1000, we will use this data to train, validate and test our future model.

Next step would be to verify that proper values are appearing in each observation, and that there were no mistakes entering or copying the data. It’s important to do so before cleaning, in order not to cast away major parts of our data sets away, due to data deviations or outliers.

**Data Preparation Phase:**

Selecting the ratio’s for Train-Test-Validate sets is important, as we will not be touching, or even looking at some of this data after the division, until the necessary moment. For now, we will retain all of our variables in order to check their importance, but in further reviews we may eliminate some if they are not correlative enough to our dependent variable.

Good housekeeping is the most important part of data preparation phase, to ensure that our model isn’t exposed to potential abnormalities we can run some tests to exclude outliers, fill-in null values, and cast aside bad data.

Important factor to this data set is the categorical value of our data will require proper encoding to accurately predict our output label – default variable. One such way to deal with non-numerical data is to utilize One-Hot Encoding, a process which categorical variables are turned into more algorithm friendly data frames.

**Modeling Phase:** At this point our data, and methodology is set, and now model creation begins. It is important to build and validate models against each other to find the most suitable one. In this project we will be utilizing Naïve Bayes and Logistic regression, we will validate each model against each other.

After creating, and training the model, assessment is necessary to see how the model performs. Main metrics to measure model accuracy is the Mean Percentage Absolute Error also known as MAPE. Other diagnostic tools include Confusion Matrix, and Root Mean Squared Error.

**Evaluation Phase:**

After running the model through our gauntlet once, we are sure to find its weaknesses. In order to correct any problematic behavior, we may tackle the data set to remove some variables, normalize our numerical data, or change the variable encoding strategy for categorical labels.

If we determine to change up our model, a plan of action is required, and testing follows. After sufficient model evolution and evaluation, we can proceed to our final step.

Deployment Phase:

The last step is to streamline the process and deploy the model to full on business use. This may include a plan to integrate this tool into daily role of our employees, a creation of tutorial and training. Gathering feedback on usage of the tool, and finally monitoring, and maintenance of the project.