Machine Learning Engineer Nanodegree

Capstone Proposal

Bank loan prediction

Abdelrahman Midhat
September 9th, 2019

Domain Background

It is important to minimize risk when it comes to finances, so modern technologies should help in fields like this.
In the past, when a client requested a loan from the bank, his data with the bank was not well utilized and not analyzed scientifically well.
But now, after the emergence of a sciences capable of helping as data science we can take advantage of the data of former customers who took loans and whether they returned or not we can analyze these data and explore good results to make a useful decision for the bank when a customer requests a new loan.

Problem Statement

For financial institutions such as banks that allow customers to take loans should reduce the risk of non-return of the loan, so it is a problem worth studying to avoid the loss of funds of institutions or banks on loans will not return, we can help them by studying the data of customers who took loans in the past and the status of these loans returned or not Let them make the right decision to approve or reject future loan requests based on their customer data.

Datasets and Inputs

The dataset from Kaggle https://www.kaggle.com/omkar5/dataset-for-bank-loan-prediction

Dataset information: The dataset has 100,000 data instances, and 19 attributes including the predictors and target variable. The 19 attributes are described as follows:

- Loan ID
- Customer ID
- Current Loan Amount
- Term
- Credit Score
- Annual Income
- Years in current job
- Home Ownership
- Purpose
- Monthly Debt
- Years of Credit History
- Months since last delinquent
- Number of Open Accounts
- Number of Credit Problems
- Current Credit Balance
- Maximum Open Credit
- Bankruptcies
- Tax Liens
- **Loan Status (Target variable)**

Solution Statement

The solution to this problem is to categorize the bank's customers into two categories based on their data previously mentioned in our data.

- Clients can accept their loan requests
- Clients must reject their loan requests

This decision should be made based on experience gained from previous customer data
This solution can be implemented by using:

- KNN
- Random Forest
- Logistic Regression

Benchmark Model

For the benchmark model, we will use the following models.

- ZeroR
- RandomForest500

Evaluation Metrics

The evaluation metric for this problem is simply the F1-Score.

Project Design

Programming Language: Python 3.6

Library: Numpy, Pandas, Matplotlib, Scikit-learn

Workflow: The general sequence of steps are as follows.

a. **Data Visualization**: Visual representation of data to find the degree of correlations between predictors and target variable and find out correlated predictors.
b. **Data Preprocessing**: Scaling and Normalization operations on data and splitting the data.
c. **Feature Engineering**: Finding relevant features.
d. **Model Selection**: Experiment with various algorithms to find out the best algorithm for this use case.
e. **Model Tuning**: Fine tune the selected algorithm to increase performance without overfitting.
f. **Testing**: Test the model on testing dataset.