**Loan Defaults**

Historically, there are three major components in a bank’s decision to make a personal loan: FICO score, income to debt ratio, and available collateral. As processing power and data analysis techniques mature, it stands to reason that there may be other factors that better predict default rates than those traditionally used. The objective of this project is to explore some of these non-traditional factors and quantify their potential predictive power. This project will consider several machine learning algorithms, discussed below, to explore these factors.

The data we are considering for this analysis comes from the Lending Club, a peer to peer lending marketplace. The marketplace works by bringing together individuals looking for loans with investors looking for returns. They do this by gathering data on those applying for loans and assigning these requests a grade based on their perceived risk from which the loan’s interest rate is derived. They have made this data available online and it presents a reasonably good data set to explore this problem, however it does present some issues. By analyzing data from a single provider we inherently bring bias into the analysis and decrease the generisability of our results. Another confounding factor is that the low rate of default (around 5.2%)

While there are many possible features to investigate, this project will focus on a select few features that have been selected based on human intuition. These algorithms may be more accurate when more, or different, features are selected, however we are focusing on a few key features that are easy for a loan officer to quantify and verify without having much impact in their workflow.

**Algorithms**

*Support Vector Machines (SVM)*

SVM is a supervised learning approach to classifying data. It takes in a set of features for each instance being tested and the associated label (in our case, whether the loan has defaulted or not) and it plots the features in space. The algorithm then finds a line that separates the two sets of points (defaults/non-defaults) while maximizing the distance from the line to the points. The result is a decision boundary that is then used to make predictions.

*Decision Tree*

Decision tree is another supervised learning approach to classifying data. This algorithm attempts to ask a series of questions to separate the data into segments using lines. The algorithm then uses those decision surfaces to classify each instance being tested according to its predicted label.

*Random Forest*

Random forest is a supervised learning approach similar to Decision Trees. In Random Forest, the Decision Trees are built from a random sample of the training data and the each branch of the tree considers a random subset of the features. This is done repeatedly and then averaged to produce the classifier.

**Data**

The features used in the analysis were:

loan\_amnt: Size, in dollars, of the requested loan

home\_ownership: Living situation of applicant. Possible values: RENT, OWN, MORTGAGE, or OTHER. Coded as 1, 2, 3, or 4 respectively.

annual\_inc: Applicant’s annual income, in dollars

fico\_range\_high: The upper boundary of the FICO range that the applicant falls under

badloan: Status of the loan, a feature defining the status of the loan as GOOD or BAD, coded as 2 or 1 respectively.

**Results**

SVM

The SVM was unable to produce generate a good classifier. The best model had an accuracy of 0.86 and an F1 score of 0.07, meaning that it was not an effective way of classifying the loans.

Decision Tree

The Decision Tree was unable to produce generate a good classifier. The best Decision Tree attempted had an accuracy of 0.74 and an F1 score of 0.15, meaning that it was not an effective way of classifying the loans.

Random Forest

The Random Forest was unable to produce generate a good classifier. The best Decision Tree attempted had an accuracy of 0.84 and an F1 score of 0.10, meaning that it was not an effective way of classifying the loans.

**Analysis**

None of the three different approaches that were attempted produced usable results. Part of this may be due to the choice of not using the entire dataset due to processing time limitations. The more likely reason is that these features are not predictive of loan default and a different set of features should be considered in future projects. Running a PCA analysis on the whole dataset would likely yield better results but is outside of the scope of this project due to run time.