For this project, I wanted to create a model that would predict whether or not a loan would default. A new business tool I used was Amazon Web Services. I used the 2007-2015 Lending Club dataset from Kaggle. Since I wanted to use features that were measured before the loan was issued, it was easy to filter out features. The dataset had 74 columns but many of those columns were measured while the loans were already in effect. A lot of the columns had too many null values to be valuable. Several columns and rows had to do with joint accounts, which made up a very small portion of the data. Figuring out the value of these joint accounts did not help my objective of predicting whether or not a loan would default.

I organized the target variable into two labels, loans that are current/fully paid and loans that are in default/charged off. The Current label had around 810,000 observations, while the Default label had around 47,000 observations. Since I wanted to maximize both the percentage of positive cases predicted and the percentage of correct positive predictions, I decided F1 would be the classification metric I would choose my model by.

The first classification method I tried was logistic regression. First, I scaled the data and set the class weights as ‘balanced’ when using LogisticRegressionCV (5 folds). I found that oversampling the Default label to make up 40% of the dataset was a better method of dealing with class imbalance. I added in more features and used Lasso to deal with collinearity and choose features. I found the optimal alpha to use by testing different values using LassoCV. I created a new model after removing features that were zeroed out by alpha.

The next classification method I tried was Random Forest. My reasoning for trying this method was that it will ignore features that do not contribute to the model and I would not have to scale the data. Similar to the Logistic Regression models I used, I found that oversampling the Default label instead of setting my class weights to ‘balanced’ resulted in a better model. However, when comparing my Random Forest models to the Logistic Regression models, the Logistic Regression models did a better job at predicting the default label. When checking different probability thresholds to maximize F1 scores, the Logistic Regression model had the highest possible F1 score, which was at a probability threshold of 0.5 anyway.

Looking at the coefficients of the Logistic Regression model that gave me the best F1 score, the coefficients made sense to me. Loans with higher interest rates were predicted to be more likely to default. If the borrower had a higher income, they were more commonly predicted to have a current or fully paid loan. One coefficient that I would like to look into further is for the member ID feature. The coefficient for this feature was much stronger relative to the coefficients of the other features.I would think that Lending Club has a scheme where they identify their more dependable borrowers with higher numbers for member IDs.

If given more time, I would like to look into using the borrowers’ job titles and their purposes for applying for the loans as features. These features were reported by each borrower and there wasn’t a concrete system for how they were formatted in the dataset. I would have also liked to spend more time adjusting the parameters on the Random Forest models I tried.