Week 3 Report

Kristina Schiffhauer, Sagar Dubey, Hewitt Trinh, Jinwoo Oh, Stefanie Montgomery

This week, we focused on predicting the probability a loan will default. We used a variety of predictive modeling techniques, including Decision Trees, Random Forests, Logistic Regression, Naive Bayes Classifier, and Multi-Layer Perceptrons, with the ultimate goal of choosing a single model we can use to predict the probability of default of a loan in the coming weeks. During the course of modeling, we experimented with varying feature sets to explore the effects of leakage. Our final models use only borrower characteristics that are known at investment time and remain static in the dataset. We notably exclude variables generated by Lending Club in our final models, which we will explain in more detail later.

Each of the modeling techniques used has varying degrees of interpretability, so we used a common set of metrics to evaluate models against each other. Batch assessments, such as AUC and Accuracy, tell us performance about the model overall. Rank-ordered assessments, such as Lift and Gain, tell us the effectiveness of our model on a top cut of records. This will be especially useful to us in later weeks in determining which loans to invest in. Finally, we created a custom metric in the form of a Cost Confusion matrix, which aims to quantify the cost of misclassification and rewards correct classification.

# Key Findings

* The most important features across interpretable models in predicting probability of default were fico score, home ownership, and debt to income ratio. High fico score is associated with a lower probability of default, as is owning a home as opposed to renting. High debt to income ratio is associated with higher probability of default.
* **We have chosen to move forward with our L1 Logistic regression model in the coming weeks**. The MLP and Random Forest models performed almost as well as Logistic Regression out of sample, but we feel that the interpretability of the Logistic Regression model gives it an extra “business” edge over the other “black box” style models.
* Models built using Lending Club (LC) variables alone (Grade or Interest Rate) have about as much predictive power as models built using only borrower characteristics. This tells us that Lending Club’s models for their own variables are as powerful in predicting default, although in a different way. It’s important to note that our best model is only 50% similar to Lending Club’s grade. Models using LC only variables have much higher non-default recall and lower non-default recall than models using borrower characteristics only. This implies a different kind of tradeoff inherent in the way LC assigned grades to loan.
* We found that the hit rate (Recall) of predicting the default loans is 54%, which may not be reassuring for a risk averse investor. Precision of predicting default is at 65.5% - a bit high because of low false positives.

# Model Summary

The below table shows a summary of how models performed on a variety of key metrics, using only borrower characteristics as features. Our chosen model, L1 Logistic Regression, is the highest performing across all metrics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Key Metrics | **Decision Tree** | **Bagging** | **Random Forest** | **L1 Logistic Regression** | **L2 Logistic Regression** | **Naive Bayes** | **MLP** |
| **Accuracy** | 62.4% | 62.7% | 63.1% | 63.7% | 63.6% | 61.6% | 63.6% |
| **AUC** | 0.66 | 0.67 | 0.68 | 0.68 | 0.68 | 0.65 | 0.68 |
| **LC Grade Similarity** | 0.42 | 0.45 | 0.46 | 0.50 | 0.50 | 0.39 | 0.49 |
| **Lift at 20%** | 1.49 | 1.51 | 1.52 | 1.54 | 1.54 | 1.38 | 1.53 |
| **Expected Return (Cost Classification)** | 324.3 | 320.4 | 327.6 | 332.1 | 330.1 | 313.4 | 330.8 |

## Decision Trees

The first 4 cuts of the trees involve decisions around the following variables: terms, fico\_range\_low, home\_ownership and dti. Interestingly, the model made the biggest cut at whether the loan term is 60 months or 36 months and decides if the loan term is 36 months, it has a higher chance to be non-default. However, as the results of diagnostic analysis last week, we know that the number of 36 months loans just simply are more available than that of 60 months loans in the data set.

Some notables rules output by decision trees model for predicting defaulted loans are:

* If FICO score is low and debt to income is high and annual income is low, you’re likely to default
* Conversely, if debtors have low debt to income ratio, they are not predicted default even if their FICO score is low
* If debtors have significantly higher FICO scores, they are not predicted to default regardless of other factors.

One interesting rule from the model:

* Debtors are predicted to default if they are renting and have high debt to income ratio.

## Random Forest and Bagging

To improve the performance of random forest and bagging, we tried different CV parameters. The cv parameters that we adjusted were 'min\_samples\_leaf' and 'n\_estimators'. Parameter change doesn’t make a significant difference in the final result, so we ended up using the same parameters that were introduced in the decision tree to measure the performance of the tree models in the same condition.

Overall, random forest outperformed bagging and decision trees. Random forest shows the highest accuracy among all trees models after removing the leakage variables. One interesting finding we gained from random forest was variable importances. As noted from the decision tree, random forest also identified the term 60 months as the most significant variable followed by fico score and dti. The below bar graph summarizes the result of top 6 variable importances from random forest.



## Logistic Regression

For a logistic regression, the key parameter to test is the number of cross-validation (CV) folds. To choose this parameter, we leveraged a function in numpy called logspace. This function, when combined with the CV parameters function in our classifier, allows us to test a number of parameters to find the optimal one in our regression. We used a minimum base of -4, up to a maximum base of 50, generating 20 samples. This function would take the log of each base in between, and determine the best CV parameter for the regression. Each of our regression models, between L1 and L2, with the various variable combinations, had different optimal CV values.

Ultimately, the logistic regression model performed best in comparison with our other models. We saw around 64% accuracy in predictions, after removing the leakage variables. The variables with the highest negative impact on probability of default were FICO score, revolving balance, and annual income. This means that the higher each of these variables were, the lower the probability of default, with FICO score (both low and high thresholds) having the greatest impact of the four. The variables that had the highest impact on probability to default were the employment length, DTI, and a 60 month term. The lower each of these variables, the higher the probability of default. Both of these findings intuitively make sense, and support the findings of our other models.

The chart below shows the features with the most positive and negative coefficient weights, that when transformed into odds ratio changes, support what is discussed above.



## Multi-Layer Perceptron (MLP)

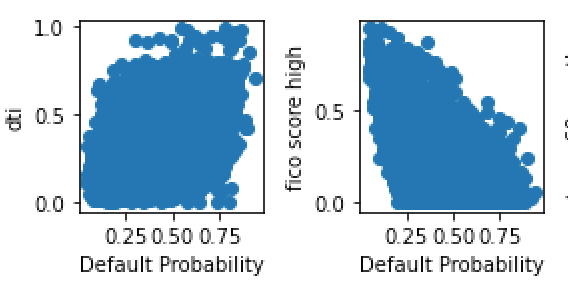
The MLP model performed almost as well as logistic regression in terms of accuracy and other batch statistics. The optimal model consisted of ‘tanh’ activation function and a single hidden layer of 100 nodes. Regardless, we chose Logistic regression as our best, to avoid the non-interpretability nature of MLP models.

## Naives Bayes

The Naive Bayes algorithm makes a strong assumption that features are conditionally independent, which often doesn’t hold true in the real world. We see that the Naive Bayes model has the lowest AUC and poor calibration. This model had the lowest overall metric scores out of sample.

# Key Relationships

We tried to see whether there are relationships between the default-probability (as predicted by logistic regression) and the significant features such as dti, fico score, term-60, revolving balance, and open account. Of these, dti and fico-score were found to have a positive and negative relationship respectively with default probability. Their scatter plots are shown below.



# Conditions and Assumptions

## Cost Confusion Matrix

In addition to batch assessment, we assess the performance of different models based on the expected returns of the models’ predictions after applying weights with confusion cost. The confusion cost was calculated based on the hybrid returns of each loan, which were detailed in last week’s report. Our confusion cost matrix is as follows:

|  |  |  |
| --- | --- | --- |
|  | Predict (Default) | Predict (Not Default) |
| Actual (Default) | 0.03 (Alternative Investment Rate ~ Bonds)  **True Positive** | -0.05 (Average Default Loan Return)  **False Negative** |
| Actual (Not Default) | 0 (Opportunity Cost)  **False Positive** | 0.092 (Average Non-default Loan Return)  **True Negative** |

-0.05 is the average return of defaulted loans and 0.092 is the average return of non-defaulted loans. In the true positive quadrant, the reward is the return from investments in stable assets like bonds at 3%. In the false negative quadrant, the cost is the opportunity cost for turning away loans, which we set it to be 0, an appropriate value relative to other gains and losses from other quadrants.

## Preventing Leakage

We experimented with varying sets of feature variables to determine the effect of leakage on our models. Here, leakage can obviously be introduced when using features related to loan performance during modeling. Total payment amount, last FICO range (both low and high), and recoveries are variables with astonishing predictive power. Using all variables (borrower characteristics, Lending Club variables, and loan performance) results in a model with accuracy of 98% and AUC of 1.0. However, the use of these variables in modeling is unreasonable as they are not known at loan investment time.

Another less obvious way that leakage might be introduced is the use of variables generated by Lending Club during the loan creation process. Examples are the Grade, Interest Rate, and Installment variables. With these included in our models, we might find that our models’ performance is strongly influenced by these variables that are created by Lending Club’s internal modeling. To test this theory, we executed our models on both Grade only and Interest Rate only. Using these two variables alone resulted in fairly accurate models, which performed about as well as models with all borrower characteristics plus lending club variables. This proved our theory true - that is, inclusion of Lending Club variables will result in models which derive much value from these features. This exercise led us to remove Lending Club-generated variables from our final models and predict default using only borrower characteristics.

## Effect of Downsampling Non-Defaulted Loans

We downsampled defaulted loans so that our dataset had a 45%/55% split of non-default vs. default, as balanced classes tend to result in stronger classification models. However, the interpretation of our model scores must be adjusted when applying our models to real-life data, as the percentage of defaulted loans will be lower. We will keep this adjustment in mind as we move to use our model in the coming weeks.