Week 4 Report

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

# Overview

This week, we applied default classification models and return regression models to test various loan selection strategies that Jasmin might employ. Jasmin must choose the number of loans to include in her portfolio, so we investigated how portfolio performance might vary with the number of loans selected. We also explored the time stability of our models, with the goal of establishing proof that our models can be trusted to choose future loans. Key findings from this week are highlighted below, and expanded upon throughout this report.

* We choose the **two-stage investment strategy** and **hybrid return calculation** as our investment framework for next week. To recap, the hybrid return calculation method is a mix of fixed time horizon method (M3) and optimistic method (M2), designed to address the shortcomings of each method. The two-stage strategy involves predicting return using a default classifier and two return regressors and is described in more detail below. We found that this strategy yields the highest return over a range of loan portfolio sizes, and should serve us well as we move into portfolio optimization next week.
* **Return regressors show slight instability over time, while default classifiers are remarkably stable.** This stability is likely because we have taken great care to ensure no leakage in our models - that is, we are limiting ourselves to only features known at the time of investment and are not updated over time. As our models are fairly time-insensitive, it follows that our strategies discussed hold across time as well.
* **The more loans Jasmin included in her portfolio, the lower her expected return would be.** This makes sense intuitively because eventually she would run out of high-return loans to select. We were able to prove this hypothesis based on empirical past loan performance (as seen in the Portfolio Size section). Our return prediction strategies show the same degradation of expected return with more loans added to the portfolio.

# Strategies Explored

We explored several different loan selection strategies. While we ultimately decided to move forward with two-stage strategy, each strategy we explored is described below:

* Random Loan Selection: The random loan strategy is the most naive of our strategies explored. It involves simply picking a random set of loans from the test set to invest in. This serves as the lower bound to benchmark all other strategies’ performance.
* Default Ranking-Based Strategy: This strategy relies on use of our default classifier model to rank loans by probability of default (in descending order). We choose the loans least likely to default.
  + We chose to use the L1 Logistic Regression classifier to predict default. As discussed last week, this model produced the best metrics overall, including highest AUC and Accuracy.
* Return-Based Strategy: This strategy relies on return prediction models developed this week. These models predict directly, regressing on borrower attributes only. We choose the loans with the highest predicted return. Note that we trained a model for each of our loan return definitions, as described in the Week 2 report.
  + Unlike previous week with classification models, we ran regression models on a full data set. The regressors that we explored were Lasso, Ridge, Linear, MLP, and Random Forest. The measurement we used to determine the best model was R2 score. MLP was the worst in terms of r2 score, but the other models showed very similar performances. Overall, random forest slightly outperformed other models.
* Two-Stage Strategy: This strategy relies on both the default classifier model and an extension of the previously mentioned return regression model. Instead of directly predicting return using a single model, we trained two models - one to predict return of loans that defaulted, and one to predict returns of loans that did not default. Again, we found that Random Forest models had the best performance and we chose to use this type of model for our separate return regressors as well. We used the default classifier model to predict the probability of default. The expected return of each loan is equal to the equation below.

*Return = E(Return|Default)\*P(Default) + E(Return|Not Default)\*(1 - P(Default)*

* + Note that it was critical in this step to adjust our default classifier for the class balance difference between the dataset used to train the model and the dataset used to test our strategy. As described last week, we downsampled the target (default) class when training classification models to the point where default and default were roughly equally represented (45% default, 55% not default). In the true dataset, defaulted loans only make up about ~17% of the dataset. We were careful to convert the probabilities that were outputs of our default classifier, which assumed balanced classes, into true probabilities representing the actual imbalance dataset. Had we not done this, our model would have been oversensitive to predicting default. The formula for adjusting the default classifier prediction is below.

A = original\_default\_prob/(train\_default\_pct/actual\_default\_pct)

B = (1 - original\_classifier\_prob)/(train\_not\_default\_percent/actual\_not\_default\_percent)

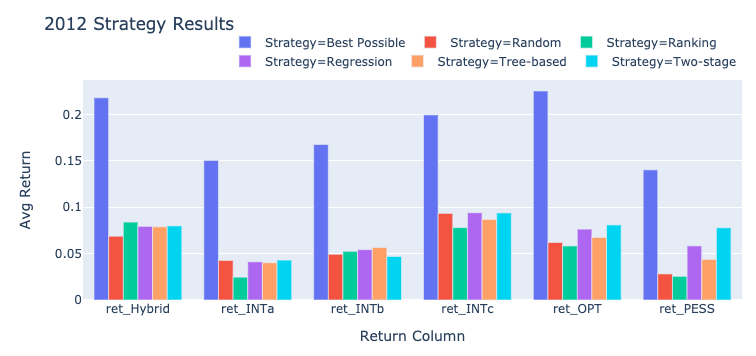
adjusted\_default\_prob = A/(A+B)

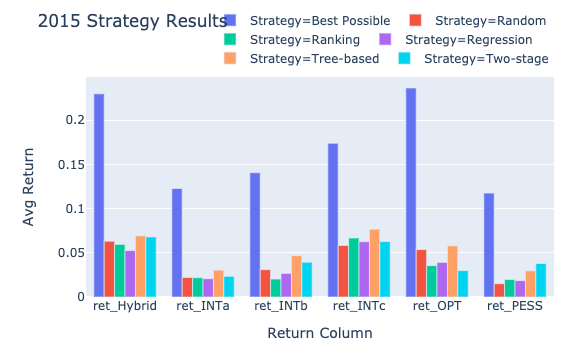
adjusted\_not\_default\_prob = B/(A+B)

For example, a loan with default probability of 29.9% from our model was converted to just 9.62% default probability in the full imbalanced dataset.

* Tree-based Strategy: We relied on the output of the decision tree model from last week’s analysis to design the set of rules for loan selection. Note that due to the fact that the number of 5-year loans is significantly smaller than that of 3-year loans, the decision tree model made the first split based on the loan term. To offset that effect, the rules we selected included only split decisions made at second nodes forward: fico\_range\_high, annual income and debt-to-income ratio (values are max-min scaled). In particular, those rules are:
  + FICO Range High > 0.363 and Annual Income > 0.448
  + FICO Range High < 0.228 and DTI < 0.358
* Best Possible: We included a “strategy” based on selecting the best possible loans to serve as an upper bound and a sanity check for our other strategies. Obviously Jasmin would not be able to employ this strategy in practice, as she would not know the best possible loans at the time of investment!

Strategy results can be seen below. Strategies were tested using models trained on data from 2012-2013. The top chart shows strategy results on the same 2012-13 dataset, while the bottom chart shows results from 2015 (models still trained on 2013 data).



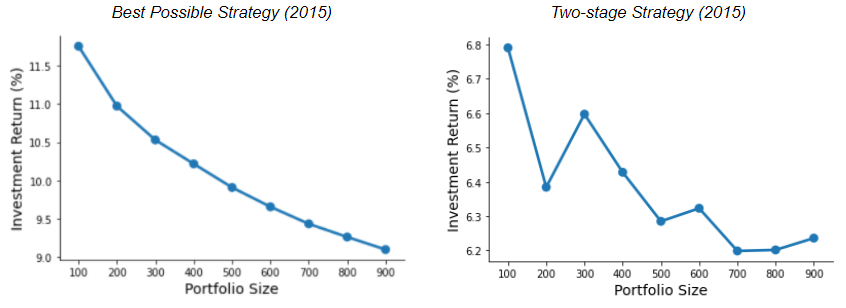


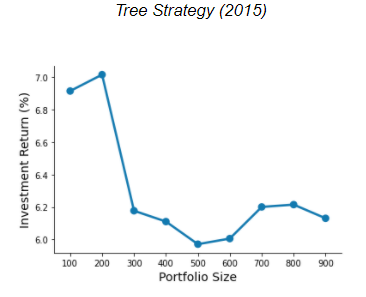
We explored these strategies across the years to ensure stability in model performance on multiple sets of loans. We found that although there are some deviations, across the two years in each respective return, the different models follow similar patterns of return in comparison to one another. This is especially the case in the hybrid return, which we have chosen to base our loan decisions off of. In both years, we see our Two-Stage model having high returns in comparison to the other models.

We can see that, generally, as our strategies become more sophisticated, they tend to perform better. The Two-Stage strategy has high performance overall with 100 loans tested on 2012 data. We see the Tree strategy slightly outperforms Two-Stage for the Hybrid return (our chosen return metric), but we have ultimately chosen to pursue the Two-Stage metric due to its higher expected return with larger numbers of loans. More on this in the Portfolio Size section.

## Portfolio size

The charts below show the percent of investment returned for 2015 for both the best possible scenario (on the left) and the two-stage strategy (on the right). We can see that even in the best possible scenario, the percent of investment return decreases as portfolio size increases. This illustrates that there are not an unlimited number of good loans to invest in, and once those have been chosen only loans with lower rates of return are left available. We wanted to ensure that our chosen strategy had the same pattern over time though we expected it to be more volatile, and we can see that same pattern reflected. It is important to note that we are operating under the assumption that we are fully, and not partially, investing in each loan. This may be oversimplifying our investments, and we will be digging deeper into optimizing both the loans to invest in and the ideal amount in each next week.



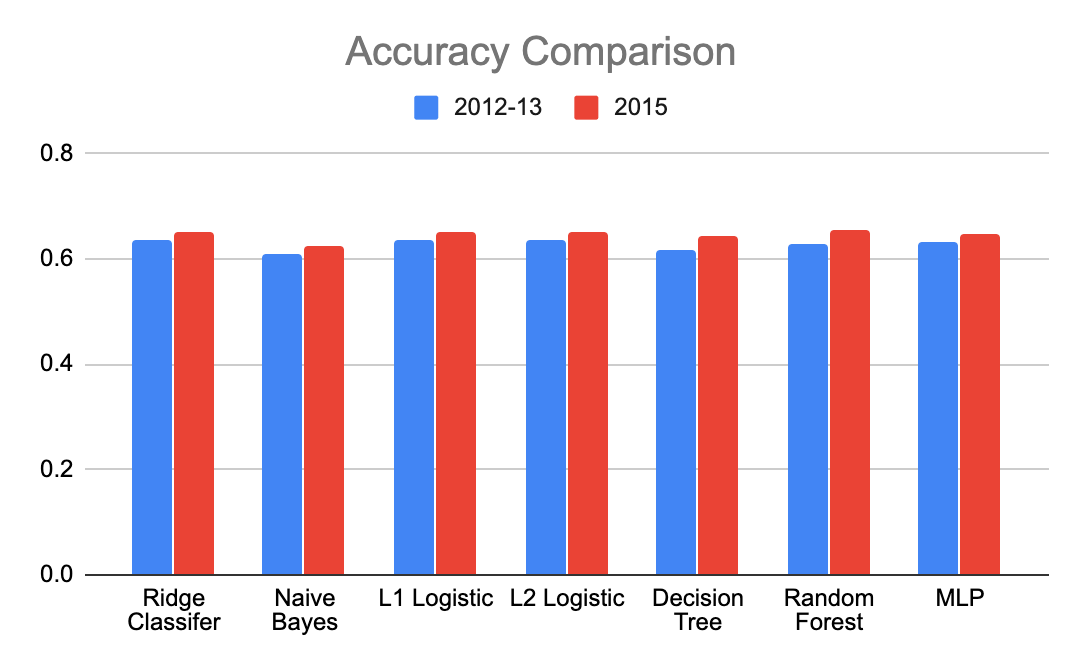
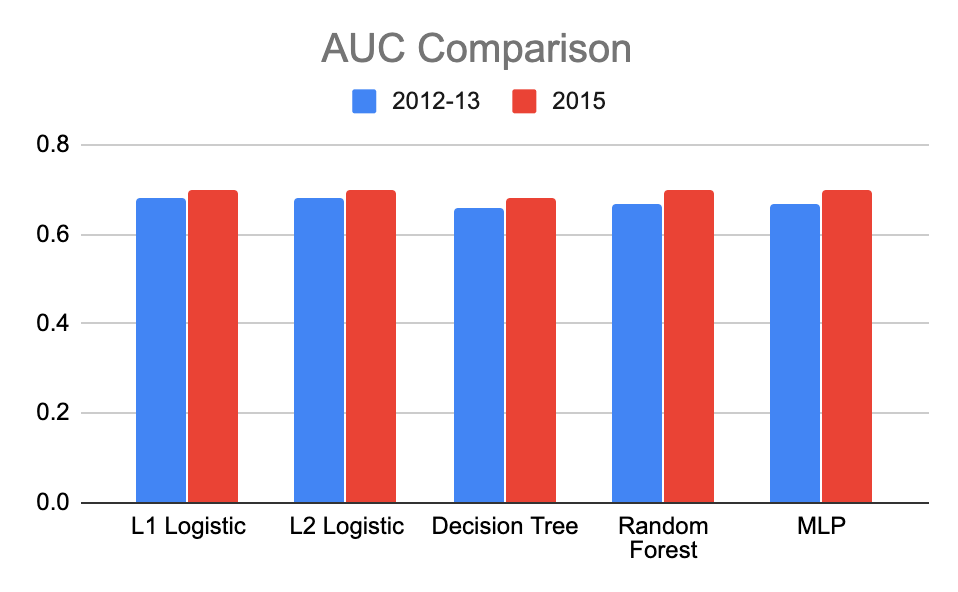


The above charts also serve to illustrate the point made earlier about the Two-stage strategy outperforming the Tree-based strategy as more loans are added to the portfolio. Though Tree-based has higher expected return with 100 and 200 loans, Two-stage yields better results with 300 or more loans. The Two-Stage model will give us more flexibility to vary our portfolio size as we move into portfolio optimization next week.

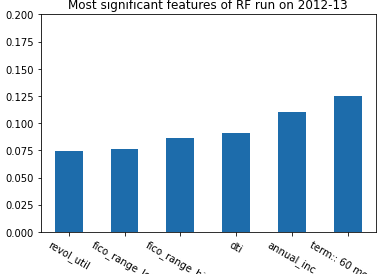
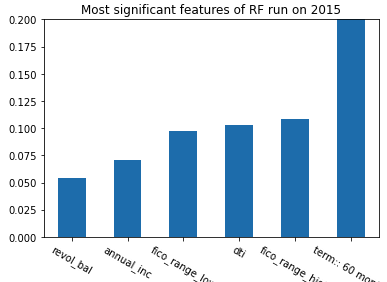
# Time Stability

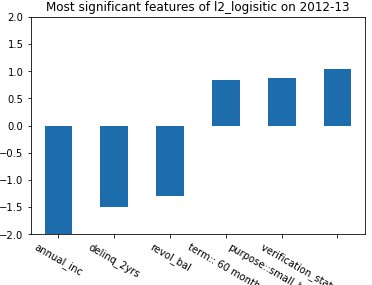
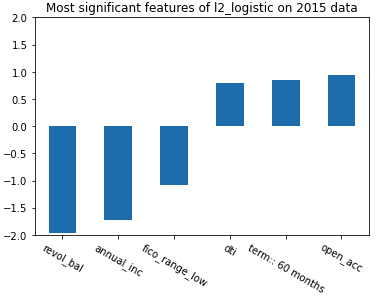
**Classifier**: We trained and tested the classifiers on 2012-2013 data. We also trained them on 2015 Q1-Q3 data, and then tested the model on 2015 Q4 data. The models trained in one period and tested on a later period performed remarkably similar in terms of accuracy and identification of top important features.

* The various models across the two time-periods performed similarly in terms of accuracy and AUC. They also performed remarkably similarly in terms of other batch metrics such as sensitivity and specificity.



* As seen for the random forest model in the figures below, three of the top 6 features - dti, annual income, revolving utilization - are the same. Three of top 6 features matched for l2-logistic regression too.





**Regressor**: Regression models were trained and tested on two different time periods to test out time stability. Unlike classifiers, the regression model has performance differences between 2012-2015 and 2015-2016 data sets. The measurement used was r2 score and it was averaged over the different return types. Although the performance gap is not significant, 2012-2013 shows better overall performance than 2015-2016 per below bar graphs. The gap is wider in ‘all returns’ compared to non-default and default.

