Final Report

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

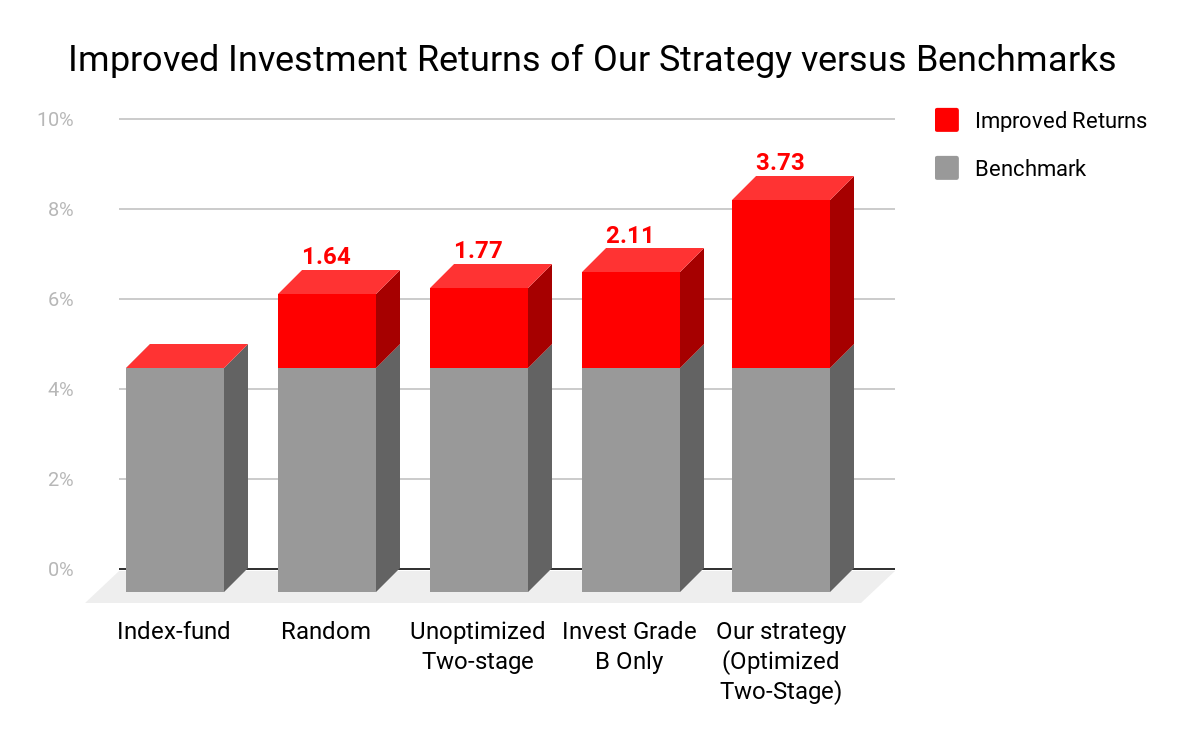
# Executive Summary

The rise of the sharing economy in recent years has brought about a new investment platform accessible to young, tech-savvy, individual investors like Jasmin Gonzales: peer-to-peer loans. What is the best investment strategy for Jasmin, a young professional proficient in analytical tools with the available data from Lending Club? The main objective of this report is to summarize the analytics journey our group has gone through to arrive at the optimal strategy, optimized two-stage, that gives her an expected return around 8.7% with a budget of $50,000 and partial investments allowed. Our conclusion is that while analytical tools such as descriptive, predictive and prescriptive models provide an edge over simplistic or naive strategies, it requires independent validations and periodic updates to maintain the desired performance level.

# Return Calculation and Benchmarks

We compared the performance of our strategy with the following benchmarks to confirm the value propositions:

* **Bonds Index Fund**: The rolling 3-year and 5-year of US Bond Index Fund is around 4 to 5% annually. We chose 5% as our first benchmark as Jasmin can easily access this type of investment with a brokerage account. Also, note that this benchmark will be the underlying assumptions of all other strategies mentioned in this report.
* **Random**: In this strategy, Jasmin is assumed to pick loans randomly and hold her investments the entire 5 years while immediately reinvesting all the payments she received from her borrowers in the above index fund. We deem this assumption to be the most realistic as other return calculations represent either overly optimistic or pessimistic assumptions about the financial markets and investors’ behaviors.
* **Unoptimized Two Stage:** This strategy simply used the outputs of our predictive models and selected the loans which have the highest predicted return. This benchmark is used to test the value of optimization in our strategy
* **Grade B Only:** In our descriptive phase, we discovered that if we can only invest in one grade of loans, we would choose grade B because it would give the highest returns. This benchmark is used to test the value of optimization and predictive models combined in our strategy.



That said, Jasmin should be aware of the volatile nature of the financial markets. Different 5-year frames in history would bring a different average return. While the relative performance among strategies is unlikely to change, the absolute return would.

# Analytics Pipeline

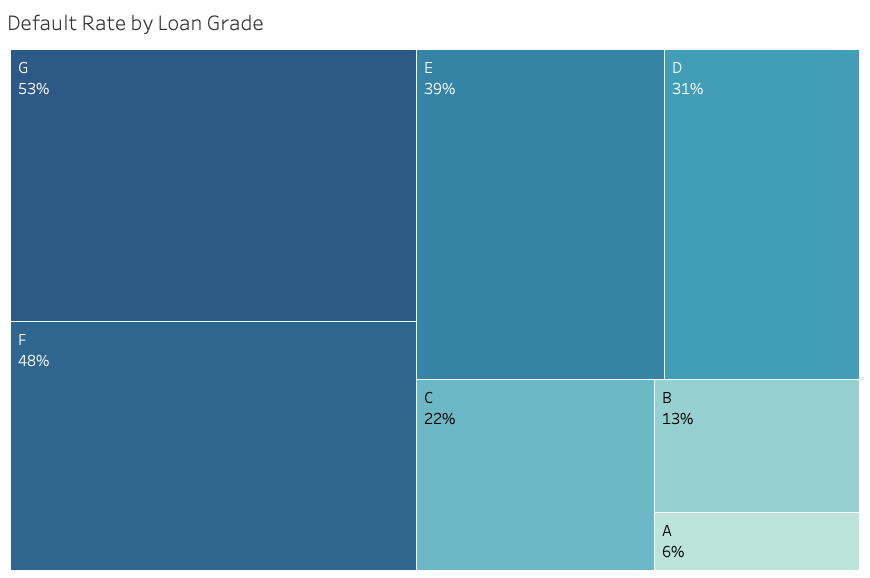
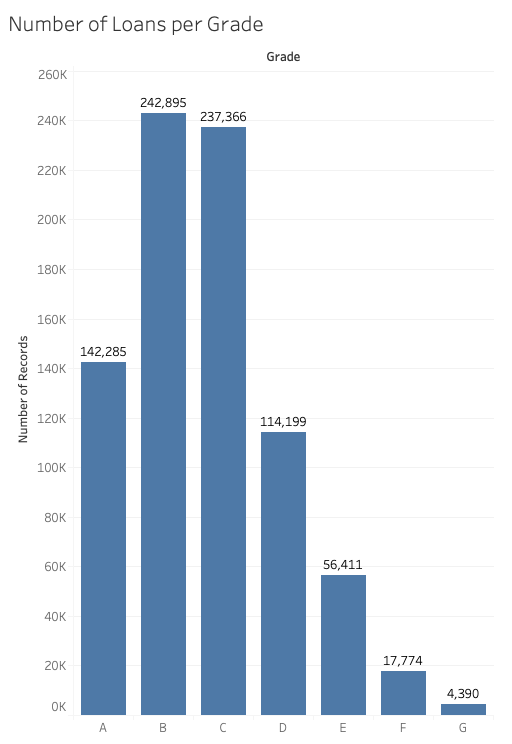
Our analytics journey began with descriptive, then predictive and finally prescriptive models. At the end, we realized that the results and findings from each stage in our pipeline can be used to perform the next or be explained by the earlier stage.

# Descriptive Analysis

Before beginning our modeling, we had to analyze, clean, and prepare our data to ensure the smoothest predictions. The data consisted of three categories of features - borrower characteristics, platform decisions, and loan performance. Borrower characteristics relate to the borrower and include variables such as annual income, FICO range, and home ownership status. Platform decisions are characteristics determined by the borrowing platform, in this case Lending Club, and include the attributes of interest rate and grade of the loan. Both borrower characteristics and platform decisions are known at the time of investment. Loan performance features include attributes such as installment amount, loan status, and total payment. These characteristics are not known at the time of initial investment, but are able to be leveraged for our models to learn which of the first two sets of features determine a quality loan.

Once we had our variables, we decided to focus on two sets of years for prior loans - 2012-2013 and 2015-2016. Though this narrowed focus allowed us to run models efficiently and get a good sense for what makes good loan predictions, it also opens us up to the risks of instability and overfitting. It is possible that these models would not perform well on a broader set of time, and we would recommend testing for additional years to confirm stability. After selecting these two groups of years for our purposes, we began the data cleaning process. Cleaning involved three key steps including removing outliers, performing log transformations on skewed variables to give them a relatively normal distribution, and normalizing all numeric variables using max-min scaling.

After cleaning our data and before beginning our predictive modeling, we wanted to get a sense of the different loan grades given by Lending Club, and how each behaves in regards to default. The charts below show the number of loans in each grade and the average default rate of each loan grade, respectively.



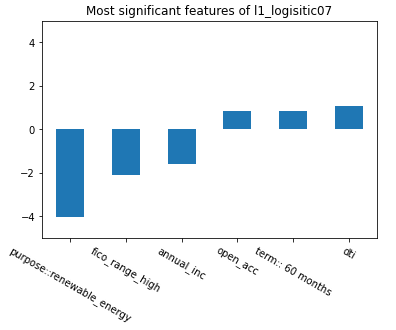
We see for grades A, B, and C, a drop in grade leads to approximately double the default rate, and the majority of loans in our sample have B and C grades. We kept these metrics in mind as we moved forward into selecting the optimal loans for Jasmin. We also performed both K means and PCA clustering with the goal of reproducing the loan grade structure, to try and get an idea of the patterns within each grade. Our K means clusters ultimately showed us that, all else equal, a longer employment length and higher income can result in a slightly boosted loan grade evaluation. PCA clustering both supported and expanded this finding, and demonstrated that FICO score, annual income, loan amount, and interest rate are the important factors to determine loan grade structures. With prepared data and deepened insight, we were able to move further down the analysis pipeline and begin the models that would ultimately predict the optimal loans to pick.

# Default Classifier

The next step in our analysis pipeline was to train a classifier model to predict the probability of default. Potential default represents a great risk to Jasmin. As we have seen in the previous section, lending club loans default at a nontrivial rate. We also know that as loan grade decreases, both the interest rate and default rate of loans increase, telling us that loans with the most potential profit are also at highest risk for default. Building a lucrative but safe portfolio hinges on being able to accurately predict default.

We explored several families of classification models, as you can see in the table below and ultimately decided to move forward with the L1 Logistic Regression model. The Multi-Layer Perceptron 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. The lift chart shown here demonstrates the power of using our default predictor to rank defaulted loans over selecting for defaulted loans at random, especially in the first ~15% of records ranked by predicted default probability.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *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 |



Logistic regression models are trained to fit coefficients for each feature that are used to output a default “score”. The higher the magnitude of the coefficient, the stronger the effect the feature has on the probability of default (as seen in the chart on the left). DTI, 60 months loan term, and open accounts positively impact the rate of default. In other words, higher values in these features correspond to a higher rate of default. The 60 month lease term variable is not entirely intuitive, and is caused primarily by class imbalance in the dataset and relative difference in behavior between 36 and 60 month loans. Even decision trees trained during this exercise made the first split on the loan term feature. Conversely, we see that annual income, fico range, and loans for renewable energy are associated with lower risk of default. It might seem strange that loans for renewable energy are so strongly associated with non default. This is because there are only 500 of such loans in our entire dataset and this small sample tends to not 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. To build a better model, 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. It is also worth noting the default classifier metrics are very stable over time. Our model results when training and testing on 2012-13 held when applying the same models to data from 2015-16.

Finally, we should highlight the fact that we chose not to use Lending Club variables such as Grade or Interest rate in our models. During training, we found that models built using only Grade or Interest Rate have about as much predictive power in terms of accuracy as models built using only borrower characteristics. This tells us that Lending Club’s models for determining their own variables are very powerful in predicting default, although in a different way than ours. Our models built using only borrower characteristics are only 50% similar to Lending Club’s grading structure. Models using Lending Club-only features have much lower default recall than our models built using borrower characteristics only, which implies a different kind of tradeoff inherent in the way that Lending Club assigns these values to their loans. We believe that our models are picking up different signals than the Lending Club variables give, which is why we chose to exclude them from our feature set.

# Return Regressor

We tried various regressors to see what model provides us with the most reliable predicted returns. Each model was trained and tested on three different types of returns: all returns, default, and non-default. To help measure the performance of each model, we use MSE as well as R². For both MSE and R², random forest shows the best performance in comparison with other models. Additionally, to verify the stability of the model, we trained and tested models on two different time periods in 2012-2013 and 2015-2016. As shown on the below graph, random forest slightly outperformed other models in terms of MSE and R². One thing to note here is that unlike default classifiers, the return regressor shows less stability over time in terms of R². Since we chose a random forest as our regressor, we further investigated variable importances to see what features contributed most to predicted returns. Top 5 features in random forests to predict returns were dti, credit history, revolving utilization, revolving balance, and annual income.

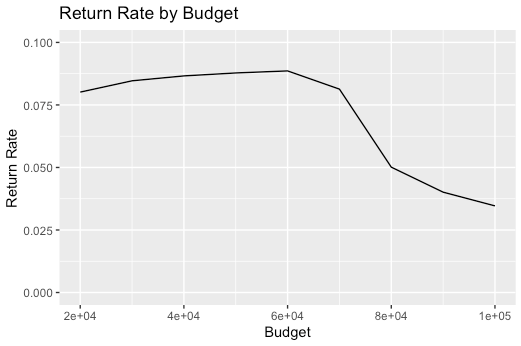
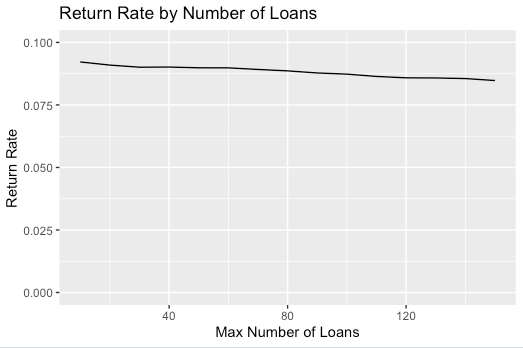
# Optimization

We combined the return classifier, default return regressor, and non-default return regressor to form the “two-stage” strategy. We calculated two-stage predicted return as expected return given default times default probability plus expected return given not default times probability of not default. We used integer programming to maximize the returns of the portfolio. Two decision variables were used. First is a binary variable to select loans or not. The other is a continuous variable to contain the partial amount invested. We set this way so that the model can allocate the partial amount to each loan. We initially set one binary variable, but this allocated the full amount to any selected loan. We realized that allowing partial investment would increase our return further because this way the allocation amount is determined considering each loan’s rate of return (higher amount for higher return, lower amount for lower return). This substantially increased our return by 3%. We set our objective to maximize (partial investment amount) \* (predicted returns from our two-stage model).

We also added constraints and created various models with different combinations of constraints. Each model is set up to select 100 loans. Models we tried are following:

* No constraint - No constraint was applied on this model except the number of loans. We had a 7% return from this model.
* Budget Constraint - The budget constraint was critical for us because this would enable us to suggest realistic investment amounts to Jasmin. The range of budget amounts we tried was 10k to 100k, and we accomplished the highest return of 8.7% with 50k of budget.
* Grade Constraint - To diversify our loan portfolio, grade constraints were applied. The constraint we applied to the model was to select at least 10 loans for each grade. This slightly lowers our return to 6.5%.
* Budget and Grade constraint - Lastly, we applied both grade and budget constraint to the model. As opposed to our expectation that this would drop our return further, the return slightly increased to 6.9%.

Once we had our optimization model up and running, we performed sensitivity analysis to ensure we were picking the best parameters to optimize our loan selection. We made the assumption that Jasmin would need to stick to a budget between $10K and $100K, and that would result in the ability to invest in around 150 loans, at a maximum. The charts below show the return rate by the maximum number of loans selected for investment and the budget, respectively.



We see that although there isn’t drastic change, the greater the number of loans selected for investment, the lower the rate of return given through our optimization. We also see that increasing the budget invested yields increasing returns until a point (around $50,000), when the return sharply decreases with increased budget. The combination of these two factors led us to conclude that a budget of $50,000 and a maximum of 90 selected loans yielded the highest reasonable rate of return.

# Portfolio Results

We are aware that selecting loans via integer/linear programming optimization is a “black box” strategy. To uncover the characteristics of the loans selected in our final portfolio, we performed Principal Component Analysis in combination with Grade breakdowns.

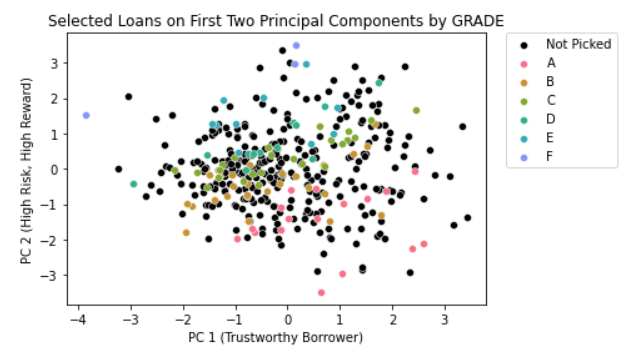
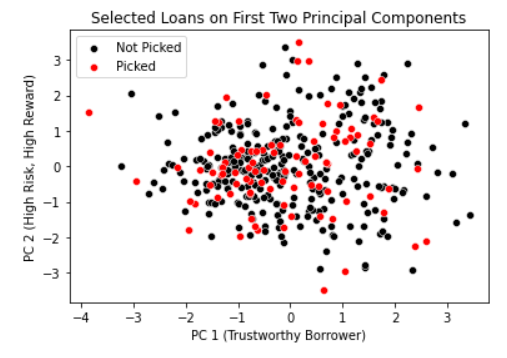
PCA was performed using the following borrower characteristics only, with loadings of the first two principal components are shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | fico\_range\_high | int\_rate | annual\_inc | loan\_amnt | emp\_length |
| PC 1 | **0.39** | -0.26 | **0.64** | **0.56** | 0.24 |
| PC 2 | **-0.54** | **0.69** | 0.20 | 0.42 | 0.10 |

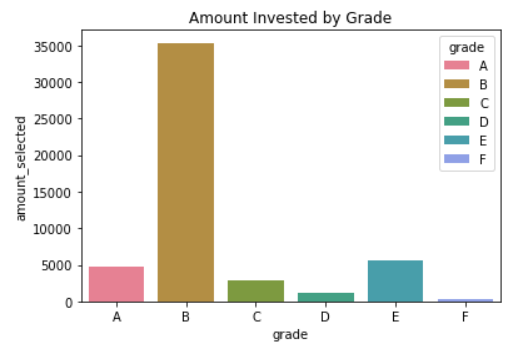
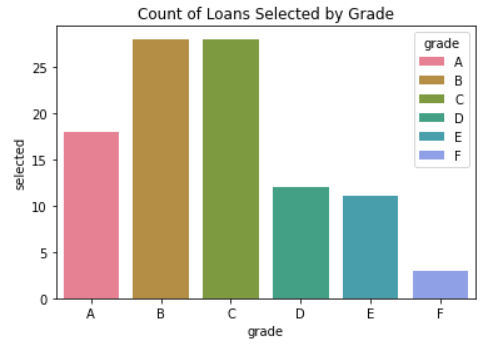
Based on our understanding of borrower characteristics and how they correspond to loan performance, we can refer to the above principal components as:

* PC 1 - “Trustworthy borrowers”: This PC is characterized by high annual income, high loan amount, and relatively high fico score. We have seen these borrowers associated with low-default, safe loans, though with comparatively low interest rates and thus lower potential returns.
* PC 2 - “High risk, high reward”: This PC is characterized by very low fico score and high interest rate. As we have seen previously, these loans are the most likely to default but can be very lucrative if the borrower does pay back the loan in full.

The below scatter plots show a sampling of loans we have selected in our portfolio plotted against the first two principal components. The left plot distinguishes selected loans (red) and not selected loans (black). This plot makes it clear that we are picking a wide variety of loans, but it is difficult to pick out exactly what kinds of loans are making up the red points. The plot on the right shows the same loans selected, but colored by Grade. We can see that the majority of our portfolio by number of loans (B and C grade loans) tends toward the center of the “risk” spectrum, which should give an investor confidence in the stability of the portfolio. We can also see that high-grade loans and low-grade loans appear to balance each other out. The A grade loans are very trustworthy and unlikely to default, contrasted with the low grade loans which are very high on the risk spectrum.



As our optimization model allows us to make partial investments in loans, we can look at our portfolio grade distribution by both number of loans selected and by amount. Looking at the number of loans selected, our portfolio favors B and C grade loans. Looking at the portfolio by amount invested tells a very different story. We can see that the vast majority of our portfolio by amount is made up of B grade loans, with A and E grade loans seeming to balance each other out in terms of risk/stability and low potential return/high potential return. The fact that our model invested in a value majority of B grade loans is encouraging, as we saw during the descriptive analytics that B grade loans historically have the highest return.



To summarize, we recommend to Jasmin a two-stage optimization strategy with which she can expect an 8.7% return on an investment of $50,000. Our analysis tells us that these are sound and safe investments that align with conventional investing wisdom. We recommend that Jasmin retrain models periodically and continuously validate model health to address any underlying data distribution changes over time. Jasmin, as an analytics professional, should have all of the necessary tools at her disposal to extract significant value out of Lending Club loans over time.