Lending Tree ROI Prediction

IE2064 Case Study#2

Jessica Starck & Loyu Liu

March 25, 2020

Section 1: Overview

Our client, Jasmin, is a young professional looking to diversify her investment portfolio by exploring peer-to-peer lending. And our task is to create a model that will predict the return on investment (ROI) for issued loans which Jasmin can utilize for a better performing and diversified portfolio. We have access to 10 years worth of Lending Club loan data, over 2 million loans, including information about the specific loan and its performance as well as the borrower(s) financial history. Our process consists of analyzing the data, cleaning the data, creating and tuning regression models, and model selection. Analyzing the data was completed in the previous Exploratory Data Analysis phase where we better understood the available attributes, the distribution of several variables and the relationship between many. The remaining steps will be described in this report and our final product results in a table providing select attributes of an issued loan and the predicted return on investment based on those details.

Section 2: Methodology

One attribute that was not included in the collected data was the return on investment for each loan, or the amount of money an investor made on each loan. We reviewed performance data of past loans of a random sampling and determined that the collected attribute, “total_pymnt_inv”, is the total amount (principal, late fees, interest, etc.) that was paid towards the funded amount. A new column, “ROI”, was added to capture the difference between the total amount paid to the investor and the amount the investor funded, the return on investment. Predicting this value for newly issued loans is the primary objective of our task and will be the “dependent” variable in our prediction models.

Because the prediction model is to be used on loans that are not yet funded, we removed all columns that would not be applicable to an un-funded loan. For example, last payment date, loan status, total payments made, etc. These values either would not yet exist or would change overtime and would not be a reasonable measure for a performance prediction. All currently “active” loans were also removed since it was unknown what the final ROI would be, as well as all variables related to joint applications only 6% of the data were for joint loans and therefore, 94% of the values were missing. Where possible and necessary, missing values were replaced (for example, a blank in employee title would be replaced with “unavailable”) or some variables were grouped into new categories to reduce the occurrence, and subsequent errors, of single values (such as grouping states into regions so there are only 5 categories that can each be represented in subsets of the data). If there were too many missing values or there was very little variance between values, rows and columns were also removed. Lastly, some variables were removed if they were highly related to other variables being utilized, such as zip code (related to region) and grade of the loan, related to sub-grade, which we chose to keep due to its greater specificity.

As a result of the data cleaning, we were left with data for 1,306,281loans and utilized 19 variables. That large of a dataset is not required to create the prediction model so we reduced it further by taking a random sample of around 6,500 loan observations. Note that large samples were tested and it was determined that the model accuracy did not improve with more observations. The last step in organizing our data is to split the data 80/20 between one data set that will be used as input to tune the models and another for the chosen model to create predictions for. The list of variables remaining in our data sets can be found in Appendix A, Figure 1.

To best determine which model is the right fit, several types were created that take on varied approaches to analyzing the data. For each model, the Root Mean Squared Error (RMSE) will be evaluated as the metric for the “fit” of the model to the data. To do this, we will find the parameters that produce the minimum RMSE value. Then, the associated RMSE Standard Deviation will be added to that value to determine an upper bound for solutions that are “close enough” to optimal. The parameters that result in the least-complex model with an RMSE value below that upper bound is the chosen solution for that regression model. This process is repeated for each model that has various parameters tested. Additionally, the R-squared values will be considered as another relative measure of fit and the simplicity and interpretability of each model type will be taken into account for model selection. The comparison of the values for each model can be found in Appendix A, Figure 2 and Figure 3, where Figure 3 includes the range of RMSE values within one standard deviation.

During this process, assumptions were made and further analysis would include validating these assumptions for a more accurate prediction model. The first assumption is the calculation of the ROI. After evaluating a sample of loans, we believe we accounted for all payments made towards the investor and an error in this calculation would alter the ROI values and if not proportional between all loans, the prediction model results would alter as well. Another assumption we made was that all data collected is accurate when in reality, there are possibilities of human error and outliers that are undesirable for predictions. Lastly, due to the large quantity of historic loan data collected and the associated attributes, we were able to reduce the data set to a reasonable size by removing all missing values, including columns that contained missing values. A

potential result is that a column with a significant relationship to the ROI was removed and therefore not included in our model.

Section 3: Results

Figures 2 and 3 in Appendix A show the comparison between the best option for each regression model created. Random forest is our recommended model for this data set because it consistently has the lowest RMSE, the highest Rsquared, and is relatively easy to interpret. Random Forest regression implements a process of creating many models internally using different subsets of variables chosen at random. Predictions are made for each internal model and then compared to choose the best by a “majority vote”. This reiterative process is also done for a range of parameters and the results of this model is the best option for each set of parameters where we then compare the RMSE values and find the simplest parameters that are sufficiently accurate. Random forest is also a good candidate because this approach reduces the correlation between the different models and with many variables in our data being highly correlated, the random sampling of variables is effective.

The results of the Random Forest model can be found in Appendix A. Figure 4 shows the top ten variables are that important to this model. Note, this does not exactly mean that these variables are the most correlated to the ROI, but it does mean that when using this model, these variables will have the most weight in the ROI predictions. Using this model and the reserved unused dataset, the chosen Random Forest model made predictions for the loans’ ROI values and the results can be found in Figure 5, 6, and 7. Figure 5 represents the distribution of ROI values for both the predicted and the true values. The difference between the true and predicted values was evaluated and the cumulative distribution of that difference is plotted in Figure 6. Both of these graphs illustrate the conservative nature of the predicted ROIs which is the preferred outcome for this application. Lastly, Figure 7 is a QQ-plot which compares the distribution of the true ROI values against that of the predicted ROI values. If the points form a relatively straight diagonal line, then the distributions are similar. As this plot shows, the results of this plot are very good and encouraging that the model is a good predictor for the loan ROIs.

Section 4: Critical Assessment

The availability of so much historical data including so many details related to each loan’s performance and the borrowers financial history proved to be a very valuable asset when completing this task. We were able to remove observations or variables with undesirable properties and still maintain a sufficient amount of data to be processed. One down side to our pre-processing efforts is the around new variables that were added to consolidate or standardize the provided data. While this effort was effective for the creation of the model, the same pre-processing will have to be conducted for any new data prior to running the prediction function. We see this as a small barrier to entry as the tidying steps are already defined and can be considered part of the overall process. We feel an improvement to our final product would be the inclusion of a probability or risk score so our client can have a better vision for the risk associated with a potential higher ROI. With further evaluation and discussions with subject matter experts, this addition would be recommended for future analysis.

Appendix A

Figure 1: Final Variables and Class Types Used

Input to regression models with ROI being the dependent variable

	data_type
loan_amnt	numeric
term	character
int_rate	numeric
installment	numeric
sub_grade	character
emp_length	character
home_ownership	character
annual_inc	numeric
verification_status	character
purpose	character
delinq_2yrs	numeric
inq_last_6mths	numeric
open_acc	numeric
pub_rec	numeric
total_acc	numeric
initial_list_status	character
diff_earliest_cred_line	numeric
loan_region	numeric
ROI	numeric

Figure 2: Model Comparison Table

model	RMSE	RMSESD	Rsquared
Ridge	5393.636	385.4761	0.0223240
Lasso	5382.291	353.6402	0.0211964
Tree	5525.053	141.0582	0.0009888
MARS	5424.105	317.6838	NaN
KNN	5771.085	457.2063	0.0067726
Random Forest	5344.273	262.0919	0.0475444
Boosting (GBM)	5393.235	310.1591	0.0208873
Bagging	5368.097	269.7586	0.0293056

model	RMSE	RMSESD	Rsquared

Figure 3: Model Comparison Plot

red horizontal line indicated the minimum RMSE

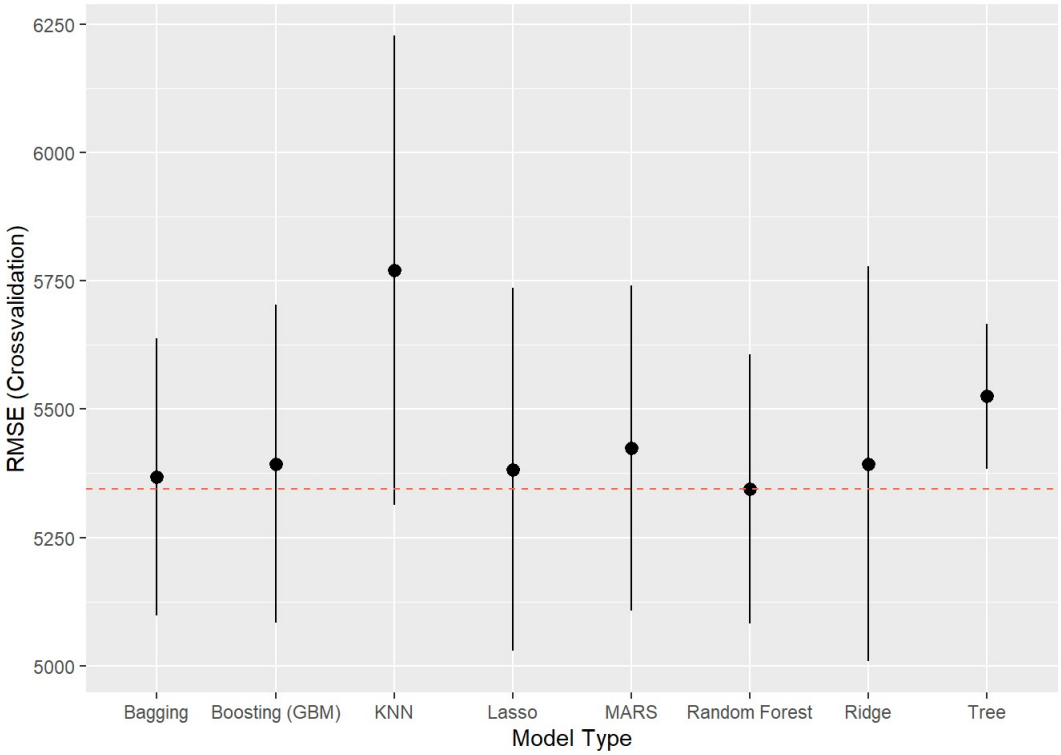
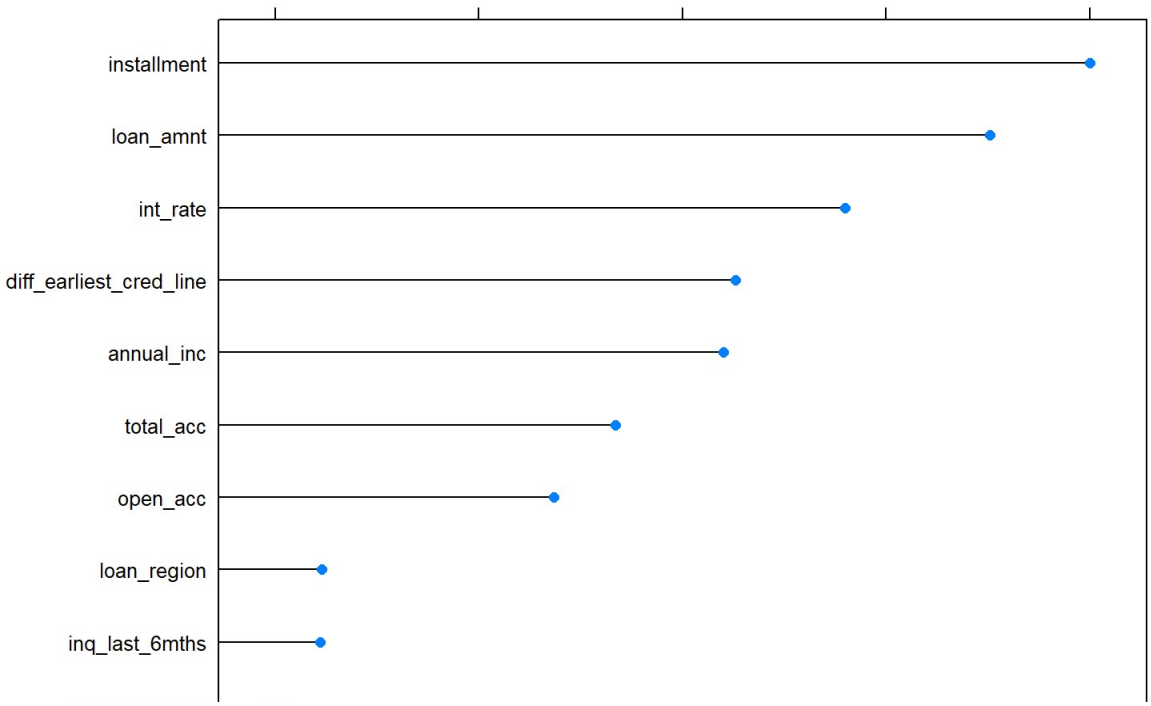


Figure 4: Top 10 Variables in Random Forest

Random Forest was the preferred model



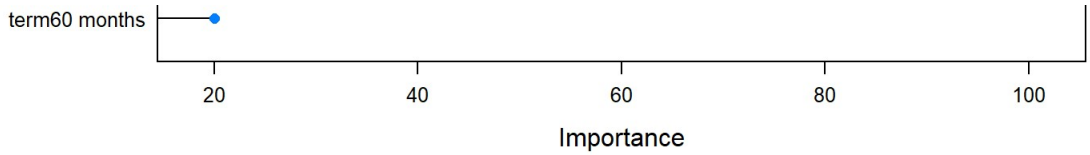


Figure 5: Densities of True and Predicted ROI

vertical lines indicate the mean for each type of ROI

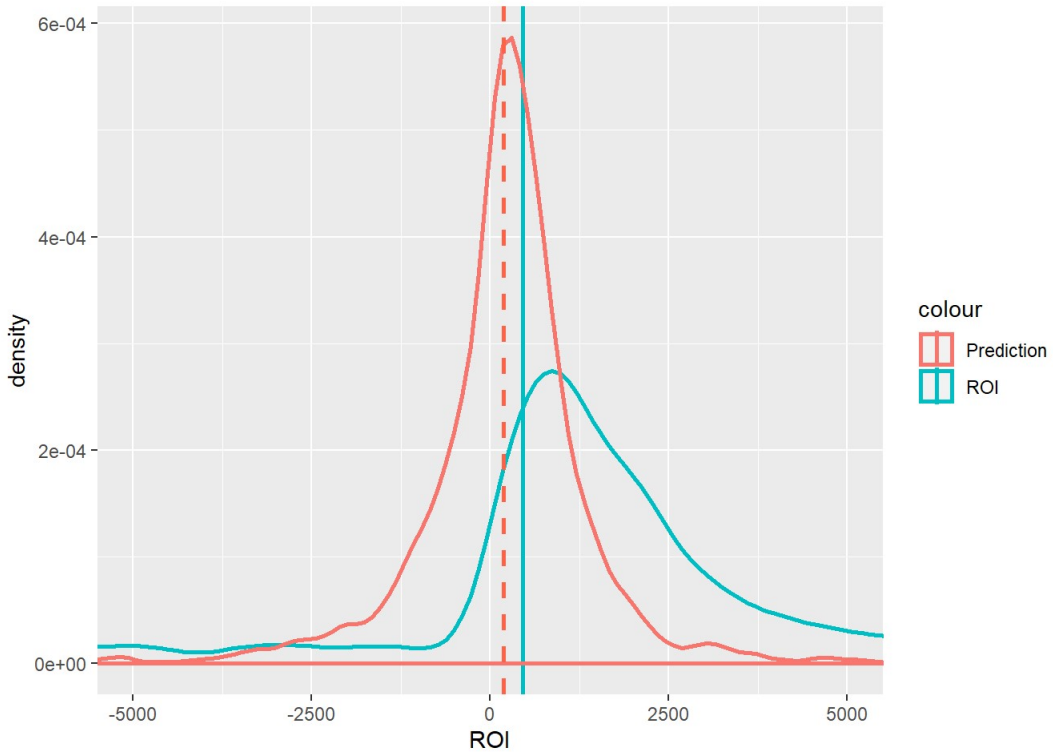


Figure 6: Cumulative Distribution of ROI Delta

positive delta indicates true value is greater than the predicted

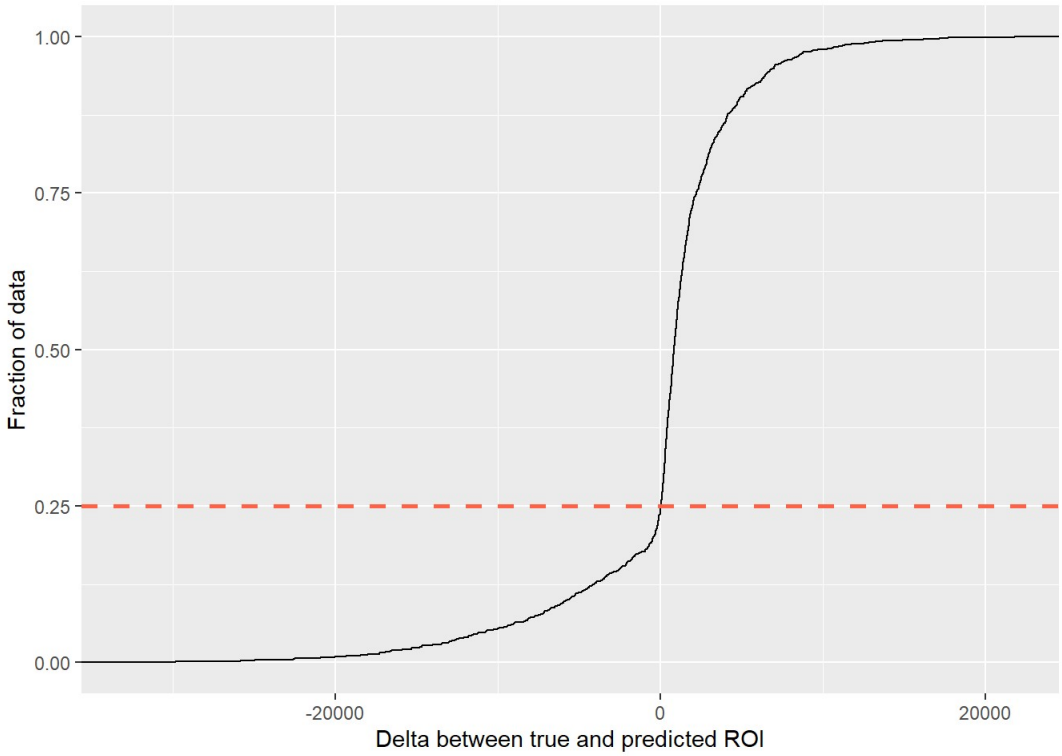


Figure 7: QQ Plot

Comparing distributions of true ROI and predicted ROI

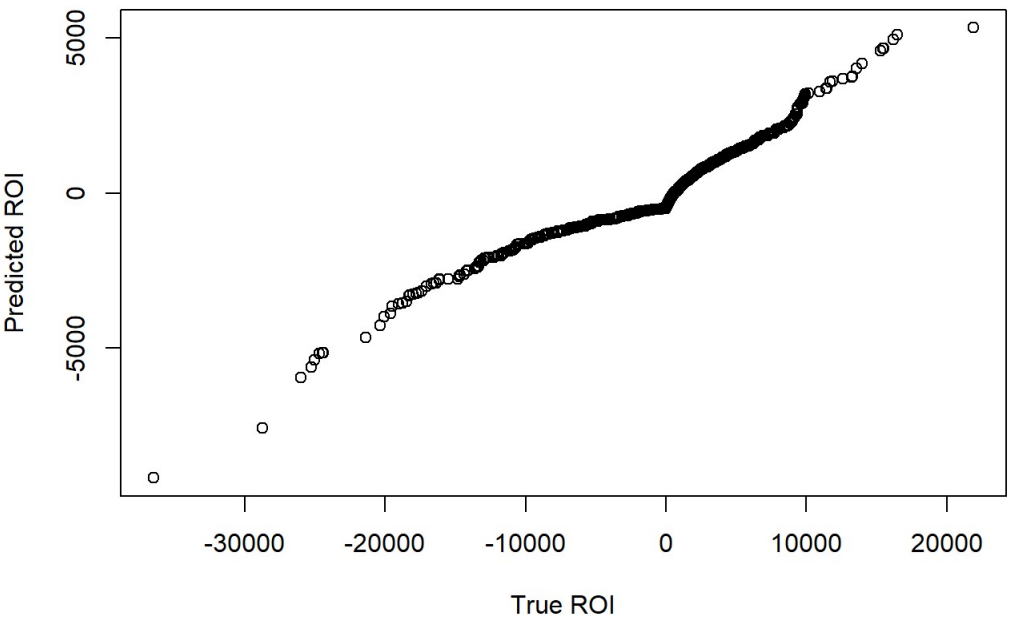


Figure 8: Example Output from Model

not all columns or rows are displayed

loan ID	predicted_ROI	loan_amnt	term	int_rate	installment	sub_grade	emp_length	home_ownership
1	831.55682	11500	36 months	6.62	353.10	A2	< 1 year	MORTGAGE
2	802.21766	14300	36 months	11.67	472.72	B4	10+ years	MORTGAGE
3	637.53684	15000	36 months	12.69	503.18	C2	2 years	MORTGAGE
4	593.08133	20000	36 months	16.99	712.96	D1	8 years	MORTGAGE
5	441.77581	10000	36 months	12.88	336.37	C2	6 years	RENT
6	51.81136	14000	60 months	13.18	319.84	C3	8 years	MORTGAGE
7	41.00634	9750	36 months	16.02	342.88	C5	10+ years	MORTGAGE
8	-153.24699	4800	36 months	6.99	148.19	A2	2 years	OWN
9	-436.65322	10000	36 months	14.46	344.02	C4	7 years	MORTGAGE
10	-612.74846	14000	36 months	15.88	491.37	C4	3 years	RENT