Paras Patel
Modeling & Valuation Associate
Interview Case Study

Credit Risk Modeling

Study Overview

Prosper Loan is a peer to peer marketplace that provides unsecured personal loans to interested borrowers and lenders looking to invest. Creditors can invest between 2,000 USD - 40,000 USD. From a lender standpoint, it is imperative to mitigate exposure to losses by understanding how a customer's financial background can influence their ability to repay the loan principal and interest. Therefore, the goal of this analysis is to assess and quantify the risk of on-going loans at the organization and any future loans.

To achieve this, a supervised machine learning algorithm is used to classify the outcome of a loan (successfully completed or defaulted). 2 methods were considered: 1) A wrapper (Linear Discriminant Analysis) for feature reduction combined with an SVM classifier and 2) A Random Forest Tree classifier. SVM (Support Vector Machine) and Random Forest are two common classifiers used in many finance applications including customer behavior, fraud, and income estimators.

The predictions of the classifier are then queried for loans labeled as 'Defaulted' (high-risk) and inputted into a regression model to estimate potential losses ($) associated with the loan.

To summarize, models in this study will answer the following questions:

1. Which borrowers at Prosper Loan are currently high- risk, based on individual background/financial features?
2. What are the potential losses on future loans based on borrower's background/financial features?
3. What are the top predictor variables that influence loan outcome and how significant are they?

Data Overview

The Prosper Loan dataset was last updated on March 11, 2014. Included with the dataset is a variable dictionary explaining all the features tracked by Prosper Loan. The dataset consists of the following structure and quality:

- 113,937 observations & 81 features

- 15 categorical variables, 39 discrete variables, 27 continuous variables
- 111,806 observations containing at least one null value
- 8 features composed of > 80% null values
- 871 duplicated observations (double entry/processed)
- 58,643 on-going loans

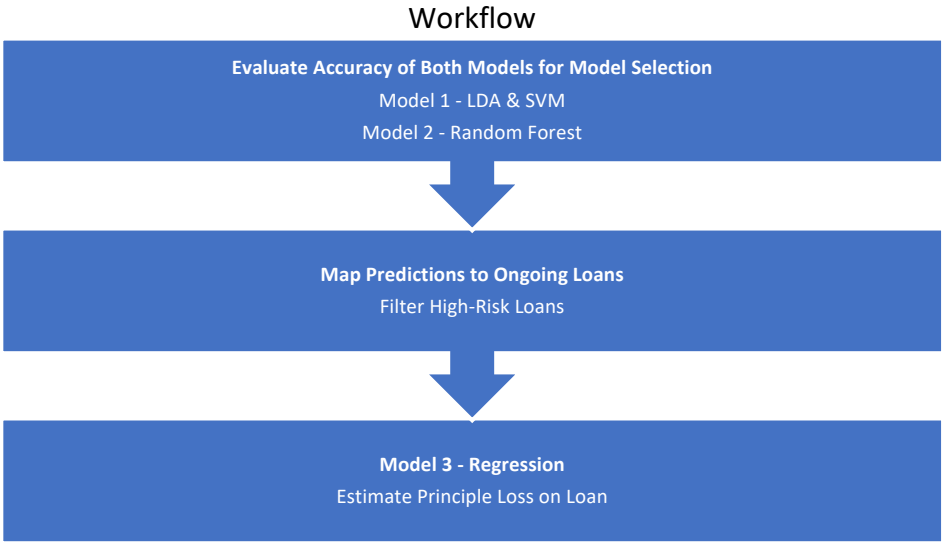
42 features were initially removed from the dataset. These features were non-discriminating, redundant, or internally labeled variables (ie. “Prosper Loan Rating”) assigned to loans by Prosper Loan.

Machine learning algorithms are highly sensitive to the quality of the input data. Models trained with bias or messy data result in inaccurate and randomized predictions. The following data cleaning techniques were used as part of data pre-processing:

- Missing Value Ratio – Any feature comprised of ‘x%’ null observations are removed from the dataset (80% used for the study)
- High Correlation Filter – Instances of collinearity between independent features identified and removed. VIF (Variance Inflation Factor) tested but not used due to high dimensionality of dataset resulting in long processing time.
- One Hot Encoding – Transforms categorical variables using a dummy encoding scheme that creates a binary column for each category level
- Duplicates removed
- Rows containing null values dropped. Performed after Missing Value Ratio to preserve number of observations

Note: Bias exists in classification model due to high skewed subgroup size in label class.

Modeling



Training for the classifiers was performed for loans with a status of “completed”, “charged off”, or “defaulted”. Due to high dimensionality and variable types of data, one hot encoding, feature scaling and dimensionality reduction are used to prepare the data for the classifiers. Two primary dimensionality reduction techniques are commonly used for predictive algorithms: PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis), the former being unsupervised, and the latter supervised.

Model 1

LDA was selected as it is optimized to separate classes because labels are known. Like other regression techniques, LDA involves computing a matrix inversion, which is inaccurate if the determinant is close to 0 (i.e. two or more variables are almost a linear combination of each other). Instances of multicollinearity were assessed and removed. The resulting components are ranked by explained variance. The explained variance of the 2 LDA components are as follows:

- LDA-1 = 65.45%
- LDA-2 = 34.55%

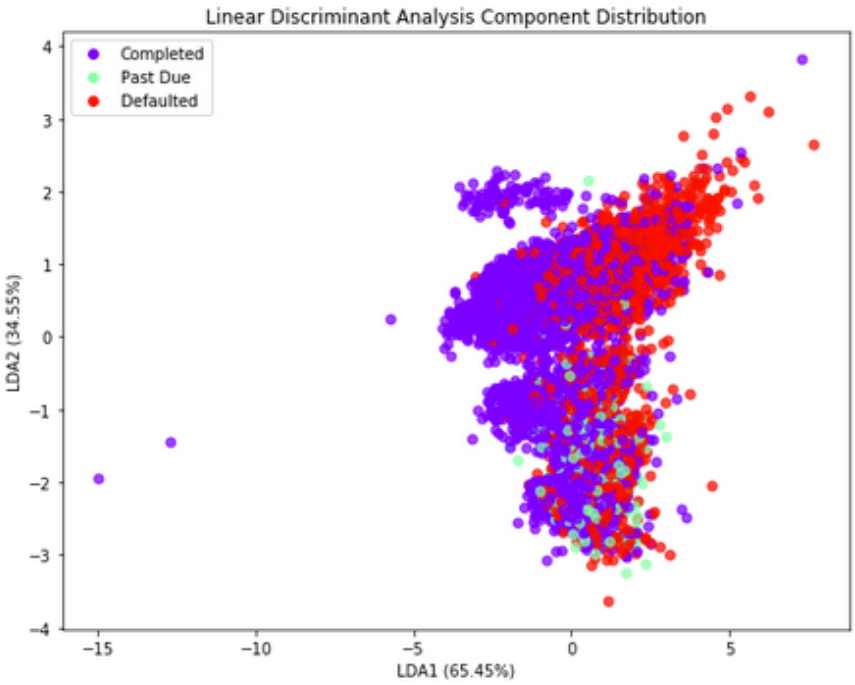


Figure 1 – LDA Distribution

The two components can be inputted to an LDA model, acting as a linear classifier, to predict new data. The LDA models prediction accuracy was **68%**. For comparison, the LDA components were inputted to an SVM classifier. An SVM model outputs an optimal separating decision surface (hyperplane) that categorizes the feature data. SVM’s are robust for both linear and

non-linear relationships. The SVM model, following cross validation, resulted in a slightly higher accuracy of **70%**. The following parameters were specified for the SVM classifier:

- kernel = 'rbf' (Gaussian kernel that calculates angular distance between 2 points)
- C = 1.0 (Regularization parameter, controls tradeoff between smooth decision boundary and classifying training points correctly)
- gamma = 'scale' (Defines influence of individual training examples)

Table 1 – SVM Confusion Matrix

Confusion Matrix - SVM Model			
Predicted Actual	Completed	Paid Off	Defaulted
Completed	6270	0	286
Paid Off	358	0	11
Defaulted	2135	0	470
Model Accuracy	70%		
Training Time	22.0 seconds		

Model 2

Random Forest is a stochastic algorithm designed to predict outcomes based on the mean of the ensemble of decision tree's the model is comprised of. The goal of each individual decision tree is to maximize information gain (function of entropy). To ensure accuracy of the model, one hot encoding was used as high level categorical features create bias in the random forest model. The resulting prediction accuracy of the classifier is **72%**. The following model parameters were specified for the algorithm:

- min _samples_split = 25 (Prevents splitting of nodes that don't have at least the specified number of samples/probabilities in them)
- n_estimators = 100 (Number of Decision Trees in the ensemble)
- Criterion = entropy (Nodes split based on entropy – comparable to results produced using gini)

Table 2 – Random Forest Confusion Matrix

Confusion Matrix - Random Forest Model			
Predicted Actual	Completed	Paid Off	Defaulted
Completed	8757	0	551
Paid Off	494	0	41
Defaulted	2765	0	1006
Model Accuracy	72%		
Training Time	6.9 seconds		

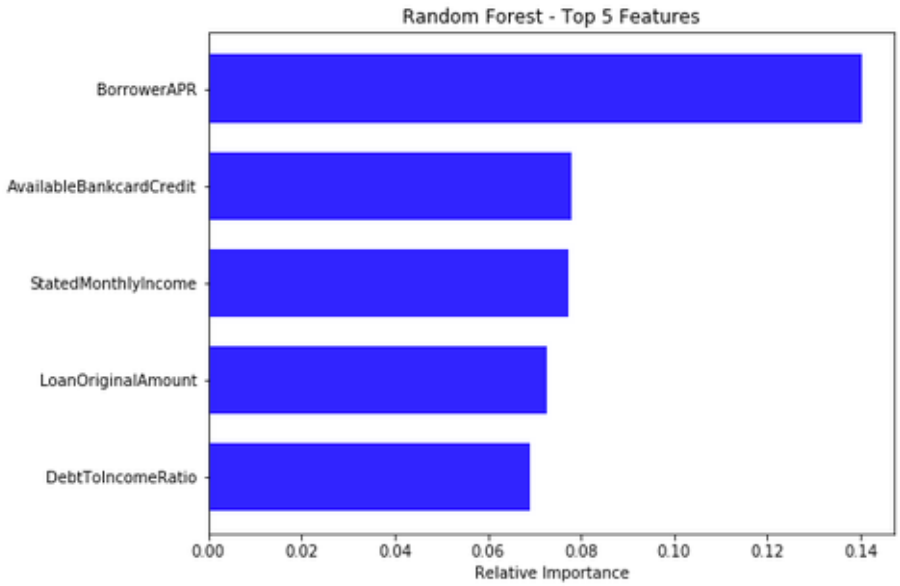


Figure 2 – Random Forest Feature Importance

Random Forest also functions as a feature selection tool. The top features, shown above by the level of importance to the predicted variable (translated to percentage), are:

1. Borrower APR
2. Available Bank Card Credit
3. Stated Monthly Income
4. Loan Original Amount
5. Debt to Income Ratio

Model 3

Random Forest was selected as the classifier of choice for the dataset. Next, all loans with a status of "Current" were queried and a prediction, using the random forest model, was made to determine the expected outcome of the loan.

Of the 50,987 loans currently on going, 2278 are labeled as "Defaulted" (i.e. high risk). This suggests that loans from Prosper Loan have an estimated **4.5% ± 1.25% default rate**, given the accuracy of the model. The loans deemed high-risk are then inputted into a regression model to predict the Gross Principle Loss of the loan.

The multi-linear regression model is evaluated by 2 metrics: R^2 and Root Mean Squared Error (RMSE). As shown below, an RMSE of $1,837 indicates one standard deviation from the regression line. With 95% confidence, any prediction from the regression model has a tolerance of $\pm 2\sigma$, resulting in an interval of $\pm$ $3,674.

Note: Loans that are "Charged Off" are considered "Defaulted".

Table 3 – Regression Model Evaluation

Regression Model Evaluation	
R-Squared	83.41%
MSE	$3,376,075.51
RMSE	$1,837.41

A final data-frame is constructed for all current loans, concatenating all the important features of the original dataset with the predictions from the models used in this study.

Listing Key	Term	Loan Status	Loan Status Predicted	Gross Principal Loss Predicted	Borrower APR	Lender Yield	Credit Score Range Lower	Stated Monthly Income	Available Bank Card Credit	Loan Original Amount	Monthly Loan Payment	Debt To Income Ratio
1027360249950330 8B223C1	36	Current	Completed	0	0.120	0.082	680	6125	10266	10000	318.93	0.18
0EF53560024827152 99901A	36	Current	Completed	0	0.125	0.087	800	2875	30754	10000	321.45	0.15
0F023589499656230C5E3E2	36	Current	Completed	0	0.246	0.199	680	9583	695	15000	563.97	0.26
0F05359734824199381F61D	60	Current	Completed	0	0.154	0.121	740	8333	86509	15000	342.37	0.36
0F0A3576754255009D63151	36	Current	Completed	0	0.310	0.261	680	2083	1929	3000	122.67	0.27
0F1035772717087366F9EA7	36	Current	Completed	0	0.239	0.192	700	3356	2181	10000	372.6	0.24
0F123545674891886D9F106	60	Current	Defaulted	10257	0.275	0.239	640	7500	363	13500	395.37	0.12
0F1C3583260311305D68F87	36	Current	Completed	0	0.131	0.093	740	5833	19129	8500	275.63	0.09
0F353575943675863D1AFC0	60	Current	Completed	0	0.117	0.085	740	10833	42204	19330	415.37	0.2

Conclusion

The goal of the study was to characterize the loans from Prosper Loan and evaluate risk associated with current loans and future loans. The assessment is based primarily from the borrower's financial background and terms of the given loan.

To accomplish this, an LDA model is used for dimensional reduction. The components of the LDA model was used for an SVM classification model. This resulted in an accuracy of 70%, with a rather long training time of 22.0 seconds. In comparison, the Random Forest model resulted in an accuracy of 72% and a much faster training time of 6.9 seconds.

The classifier was then used to query current on-going loans labeled as high-risk. These loans were inputted into a regression model that estimated the Gross Principle Loss associated with each loan.

The models used in this study can significantly reduce negative exposure for a lender seeking to lend money to a borrower or at the least, evaluate risk based on a few predictive features identified from the models.

Source

www.github.com/pjpatel012