Using Predictive Analytics to Design Investment Strategies for Fintech Lending

Constantin Romanescu

https://github.com/cromanes/Ryerson-Capstone

Table of Contents

Introduction

With the advent of on-line technologies, Peer-to-Peer (P2P) lending has become an important alternative to traditional bank lending. In a nutshell, a P2P company is a credit marketplace that matches borrowers with lenders (investors). The first such company was Zopa, founded in 2005 in UK. Today, Fintech companies are present in many countries, with US and China being the major players.

Lending Club, a Fintech company founded in 2007 and headquartered in San Francisco, USA, is the major on-line lending platform with more than 41 billion USD loan issuance (as of September 30, 2018). In 2015 it became the first P2P company listed on Stock exchange. Lending Club (LC) offers small (1,000 – 40,000 USD) unsecured loans for 36 or 60 month periods. Based on the credit worthiness of the customers, LC assigns each loan a grade (A – G) and a subgrade (1 – 5), with A1 being the best classification. The loan grade is used to assign the interest rate. The loans are then listed and investors can commit to cover a number of notes (each 25 USD in value). If a loan is not fully funded in 30 days, the borrower can opt for a smaller funded amount. Each loan must be repaid in monthly equal installments throughout its maturity period.

Along with other P2P companies, Lending Club makes their data publicly available.[1] This represents an excellent opportunity for investors to make better informed decisions in picking up investments. Also, researchers can use the data to propose and test analytics models to study loan defaulting.

In the remainder of this work we will use the Lending Club data to design a data analytics product that will help investors pick the loans with the best annualized rates of return. The process follow the steps outlined by the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000) and includes data ingestion and processing, and machine learning techniques (binomial and multinomial classification, regression).

[1] https://www.lendingclub.com/info/download-data.action

Literature Review

1. *Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending (Berger & Gleisner, 2009)*

The authors look at the dynamics of the interaction between the lenders and the borrowers in both P2P and traditional banking transactions in the early days of social lending. Using statistical analysis on a dataset provided by Prosper.com, the paper concludes that intermediaries create value by reducing the information asymmetry between lenders and borrowers. As a result, borrowers seen as risky customers might find it easier to access credit in an electronic market.

2. *P2P Lending Survey: Platforms, Recent Advances and Prospects (Zhao et al., 2017)*

This paper provides succinct information on the main P2P players around the world and discusses issues related to P2P from a social and economic perspective. It provides good explanations for the technical terms used in the industry and the decision making process.

3. *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform (Jagtiani & Lemieux, 2018)*

As the title suggests, this working paper looks at the role of alternative data in the loan grade assignment at Lending Club. Using regression analysis on the data from 2017 – 2015, the authors report that the correlation between the loan grades and the FICO scores declined from about 80% to only 35%. Therefore, alternative data has been increasingly used by Fintech companies. While the exact recipe for loan grade calculations is unknown, alternative data can include: utility payments, insurance claims, cell phone and internet use, bank account transfers, etc. Less reliance on FICO scores could mean that some sub-prime borrowers will get easier access to credit.

4. *Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending (Emekter, Tu, Jirasakuldech, & Lu, 2015)*

The paper presents an extensive statistical study on loans provided by LC (2007 – 2012). The authors use logistic regression to identify markers for loan default, and found that loan grade, debt-to-income ratio, FICO score, and revolving line play an important role in loan defaults. A comparison between the LC customers and the average US consumers revealed that

high FICO score and high income borrowers don't use P2P services, therefore suggesting targeting this group of potential customers.

5. *Risk Assessment in Social Lending via Random Forests (Malekipirbazari & Aksakalli, 2015)*

Malekipirbazari and Aksakalli used a number of binary classification techniques (logistic regression, KNN, SVM, RF) and false positive penalty error (5:1) using a LC dataset. In the end they proposed an RF classification model that is superior to LC internal rating and FICO scores in identifying good borrowers; however, they failed to significantly enhance the capability of their method to distinguish default borrowers. Also, provides a refresher on classification algorithms.

6. *Determinants of Default in P2P Lending (Serrano-Cinca & Gutiérrez-Nieto, 2016)*

Serrano-Cinca and Gutiérrez-Nieto proposed the use of profit scoring instead of credit scoring for ranking transactions in P2P lending using data from Lending Club. They used a well-known financial ratio, internal investment ratio (IRR), to gain insights on the expected profitability of investing in P2P loans. Their approach used Multivariate linear regression and the exhaustive chi-square automatic interaction detection (CHAID). The proposed decision rules based on CHAID achieved promising IRR, but a cost-sensitive model remains to be developed.

7. *Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Credit (Polena & Regner, 2018)*

This paper is based on the first author's master thesis published in 2015 and is somewhat dated. It uses logistic regression on data from the Lending Club (January 2009 to December 2012) to identify determinants of loan default. They found a number of features that work on overall data (all loan grades) and features that work only for predicting default for specific loan grades. The first category of features include: annual income, debt-to-income ratio, loan purpose, and Inquiries in the past 6 months.

8. *Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending (Li, Li, Yao, & Wen, 2019)*

The authors look at a LC dataset from the perspective of revenue loss and use multinomial logistic regression to predict loans that fully paid, charged off, and, prepaid (lost revenue from

interest payments). In addition to the data provided by LC, they considered, also, macroeconomic factors such as: GDP growth rate, Federal funds, and, Bankruptcy filings. The overall accuracy was ~75%. The highest accuracy value was obtained for prepaid loans (~87%). These findings could serve as benchmarks for our models.

9. *Credit Risk Prediction in an Imbalanced Social Lending Environment (Namvar, Siami, Rabhi, & Naderpour, 2018)*

The authors approach the problem of loan default through the lens of imbalance data. They used three different binary classification techniques: Random Forests, Linear Discriminant Analysis, and Logistic Regression and one of under-sampling, over-sampling or hybrid approach to handle imbalancing on LC data. In addition to the standard metrics (AUC, accuracy, sensitivity, specificity) they also used the G-mean. They conclude that random forest and random under-sampling gives the best strategy. However, it seems that many combinations give similar results. Using cross-validation would render many combinations statistically similar.

10. *Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science (Cohen, Guetta, Jiao, & Provost, 2018)*

The goal of this paper is "*to present a comprehensive case study based on a real-world application that can be used in the context of a data science course*". It outlines the main steps in a data science project. Use the LC data to run binary classifications (Naïve-Bayes, regularized logistic regression, decision trees, random forests, MLP), develop investment strategies based on return on investments, and linear optimization for portfolio optimization. Not a research paper.

11. *A new aspect on P2P online lending default prediction using meta-level phone usage data in China (Ma, Zhao, Zhou, & Liu, 2018)*

In this paper, Ma et al. approach the information asymmetry problem by using meta-data from phone usage. They used AdaBoost to make predictions on loan defaults and claim that phone and apps usage patterns are strong predictors for an individual loan default. The results seem exciting as they point to endless limits of data analytics.

12. *Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending (Jiang, Wang, Wang, & Ding, 2018)*

In this research paper the authors combined predictors used by traditional banking institutions with features extracted from the text description of the loan. They used LDA to extract credit-related topics and built loan default predictive models using four classification methods (logistic regression, Naive Bayes, SVM, and RF). They tested the models using data from Chinese P2P platform. They acknowledge the study's deficiencies and propose expanding the work to other sources of unstructured data.

13. *Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending (Xia, Liu, & Liu, 2017)*

In this math-heavy paper the authors use a cost-sensitive boosted tree loan evaluation model that combines cost-sensitive learning and extreme gradient boosting to predict the return and the risk of the loans using data from LC and We.com (China). In addition to it, the paper presents an excellent review of existing literature models. Definitely, one of best papers on the subject.

14. *Financial return crowdfunding: literature review and bibliometric analysis (Martinez-Climent, Zorio-Grima, & Ribeiro-Soriano, 2018)*

As stated in the title, this paper reviews the literature on P2P and Equity Crowdfunding (EC) that was indexed by Web of Science. While it does not provide technical insights into the problem, it nicely summarizes the literature by time and country.

Dataset

The data used in the study was provided by Lending Club and covers the 2007 – 2018 period.

The description of the features was provided by Lending Club.[2]

The dataset consists of 2,004,062 observation and 145 features (~ 1.4 GB). An important note is that 6 features from the loan description are missing from our data. These features contain ranges of FICO scores for the primary and secondary applicants. Although these are important features in predicting the payment of a loan, they are expected to be correlated with the Lending Club rating and the interest rate assigned to each loan. The primary target variable is *loan_status*.

A summary of the target variable is given in table 1.

[2] https://resources.lendingclub.com/LCDataDictionary.xlsx

Table 1. Count and percentage summaries of the loan_status levels.

loan_status	Count	Percentage
Fully Paid	873920	43.607
Current	865418	43.183
Charged Off	220712	11.013
Late (31-120 days)	21092	1.052
In Grace Period	14370	0.717
Late (16-30 days)	5766	0.288
Does not meet the credit policy. Status: Fully Paid	1988	0.099
Does not meet the credit policy. Status: Charged Off	761	0.038
Default	35	0.002

There are two main issues with the dataset that need to be addressed:

- filter the observations that contain relevant data;
- find and remove features that leak information about the target variable.

Approach

The analytics pipeline used in this project follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach. (Chapman, et al., 2000) CRISP-DM is a six-step iterative and adaptive process, as shown in Figure 1.(Larose & Larose, 2015)

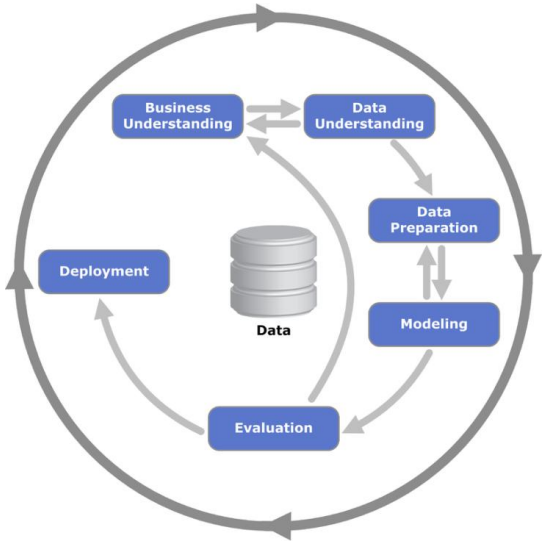

Figure 1. CRISP-DM pipeline diagram[3]

[3] https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Step 1: Business/ Research understanding

The details pertaining to this phase were discussed in the introduction and the literature review section. The goal of the project is to provide a machine learning tool for identifying the best performing loans (and avoid defaulting loans) using the Lending Club data.

Step 2: Data understanding – Gain insights from data

Details on the work performed in steps 2 – 3 are provided in the accompanying files on GitHub.[4] The work includes:

- Download data;
- Exploratory data analysis; Familiarize with the data; Corroborate the insights with the literature review; Remove features added to the dataset after the loan was granted;
- Evaluate the quality of the data: Remove features with (near) zero variance, many missing observations (50% threshold), or with too many levels (50) for categorical data;
- Look at alternative ways to measure a loan's success;
- Introduce features containing socio-economic data.

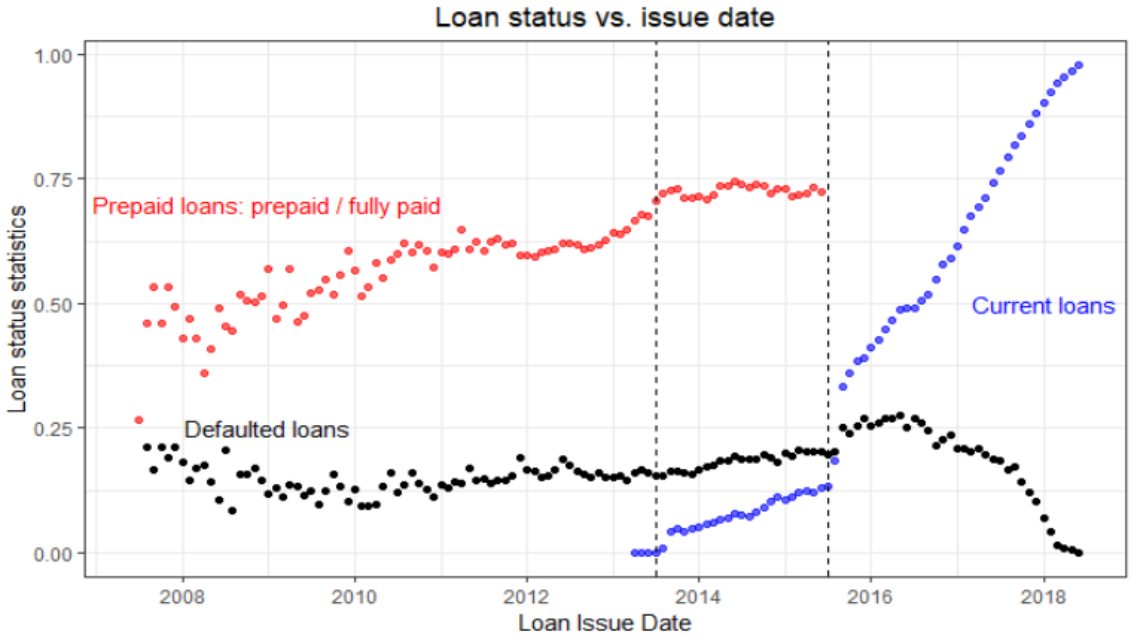

Figure 2. Loan status as a function of loan issue date

[4] github.com/cromanes/Ryerson-Capstone

The traces in Figure 2 show the evolution of the main loan status categories for Lending Club data reported on June 30, 2018. The regions to the left of the vertical bars represent matured loans with a 60 month and a 36 month term, respectively. We notice a few trends:

- the default rate, charged_off / (charged_off + fully_paid) for matured loans increased slightly over time;
- the 36-month term loans make the majority of loans and have smaller default rates compared to the 60-month term loans;
- the prepayment rate, prepaid/ fully_paid, increased from 50% to 75%;
- loans issued after June 2015 have to be treated separately.

To better understand the dynamics of loan (re)payment we introduced a new variable, months-on-the-book (MOB) defined as the time in months between the last payment and the loan issue date for the loans that reached their maturity period. In Figure 3 we show the count bar plots for the prepaid loans and the default loans, for the 36-months term loans.

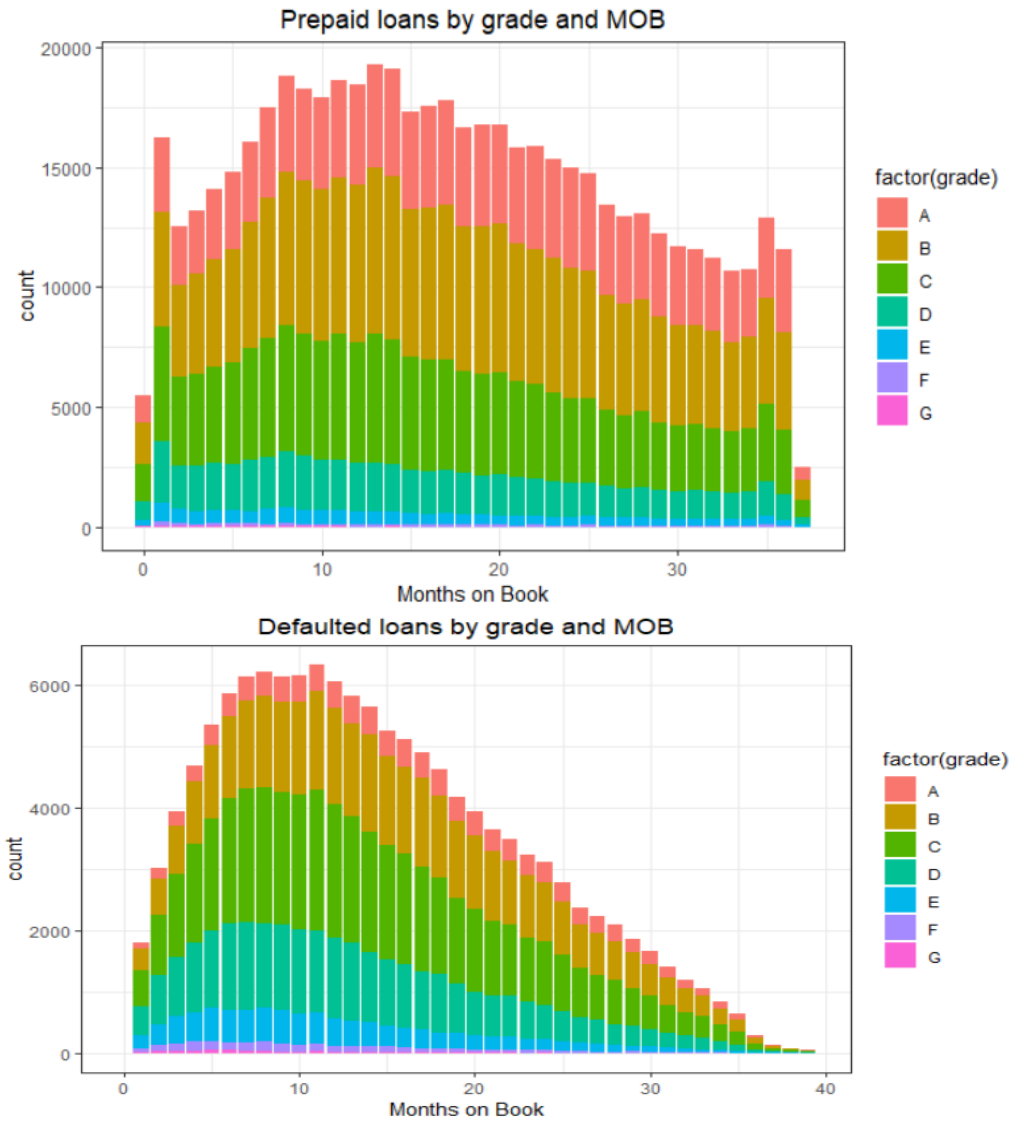

Figure 3. Prepaid and defaulted loan counts as a function of number of months on the book.

On average, a borrower defaults on a 36-month loan after 14.5 months or prepays the loan in 17.3 months.

We also analyzed the default and the prepay distribution as a function of MOB and loan grade or the amount borrowed and did not find significant differences between payment patterns. The analysis is in close agreement with the literature data.(Li, et al., 2019)

Given that there is no charge for prepaying a loan and prepayment leads to decrease in projected revenue for the investors it would be interesting to look also at the multinomial classification of the loan status.

At this point, we decided to carry on the analysis using the matured 36-months loans issued after January 2012.

Following the suggestion made by (Cohen, et al., 2018) we decided to introduce a new dependent variable, the annualized return on investment, ROI, which is defined as:

$$ROI = \frac{total\_payment - loan\_amount}{loan\_amount} \times \frac{12}{36}$$

This is a rather *pessimistic* approach, as it assumes that once a payment is received it sits for the remainder of the 3-year term. Most of the defaulted loans result in a negative ROI. We also calculated an '*optimistic'* ROI and a '*realistic'* ROI using the formulas suggested by Cohen et al.

After introducing new independent variables: unemployment rate at state level, average credit card rate, S&P 500 monthly average index, and the T-bills monthly averages, we are left with a dataset containing 396,682 observations and 29 predictive features. We split the data into train/ validation/ test (60:20:20) datasets. In November 2018, Lending Club published its quarterly update which gave us 3 months of new matured loans data (from July – September 2015). This data provided us with a second test set.

Step 3: Data preparation

In this step we report data manipulation steps based on the train data. Although the creation of new features belongs to this step, it was moved to the previous step in an attempt to simplify data workflow.

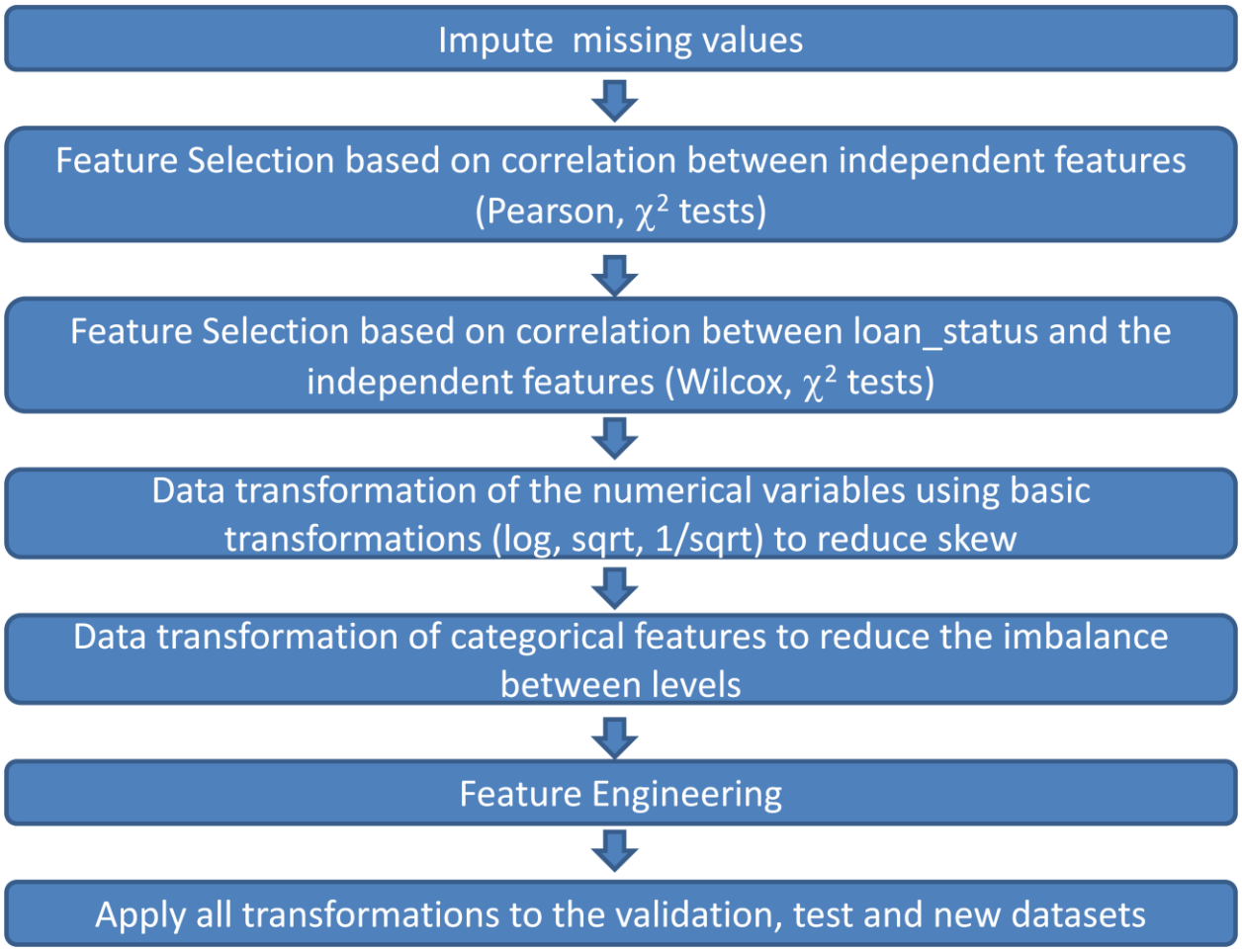

Figure 4. Data preparation workflow

A summary of the numerical data after the feature selection step is shown in Table 2.

Table 2. Numeric data summary

	Min.Val	Mean	Median	Max.Val	StDev	Skew	Wilcox
loan_amnt	1000.0	12561.3	10000.0	35000.0	7668.4	1.0	TRUE
int_rate	5.3	12.4	12.3	27.9	3.9	0.4	TRUE
annual_inc	3000.0	71784.7	60000.0	7500000.0	60334.5	36.9	TRUE
dti	0.0	17.5	17.0	40.0	8.1	0.3	TRUE
total_acc	2.0	24.8	23.0	162.0	11.7	0.9	TRUE
pub_rec	0.0	0.2	0.0	40.0	0.6	6.9	TRUE
revol_bal	0.0	15373.4	10790.0	2904836.0	22026.4	31.9	TRUE
revol_util	0.0	54.8	55.5	892.3	23.4	0.0	TRUE
delinq_2yrs	0.0	0.3	0.0	24.0	0.9	5.4	TRUE
Credit_history	36.0	192.0	174.0	780.0	89.8	1.1	TRUE

X3_Yr_avg	0.3	0.9	0.9	1.3	0.2	-0.7	TRUE
CreditCardAvg	11.8	11.9	11.9	12.4	0.1	1.5	TRUE
Unemp_rate	2.9	6.6	6.4	12.3	1.4	0.4	TRUE

Note the scale difference and skew of the variables. An example of highly skewed data is shown in the figure below:

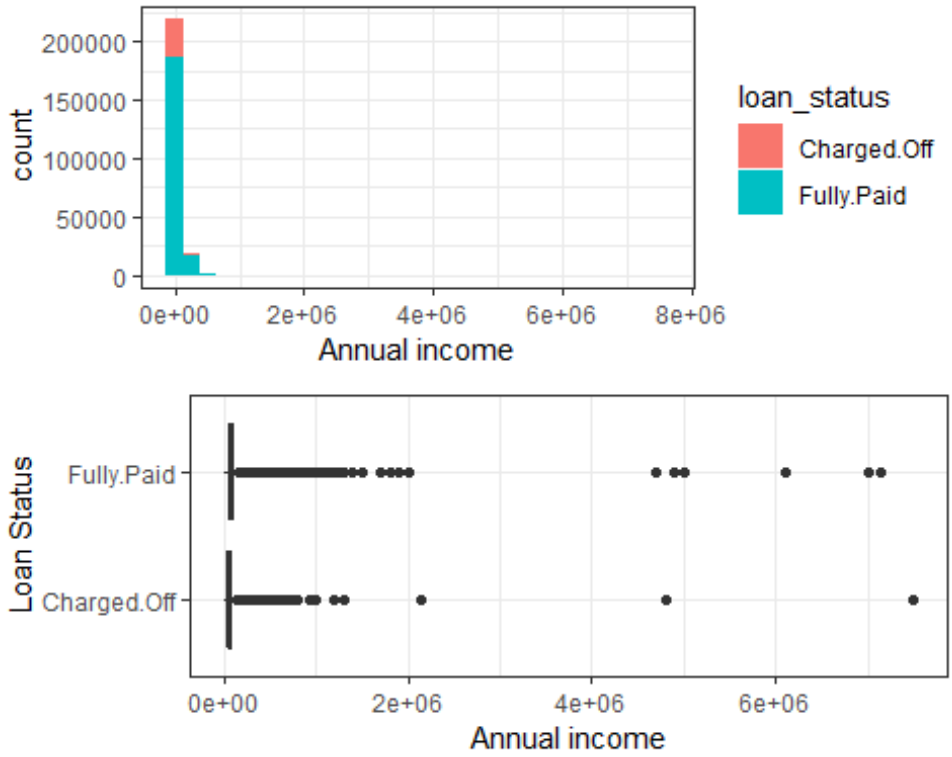

Figure 5. Conditional histogram and boxplots of annual income

After applying the log transformation, the skew is reduced to 0.225. Detailed data analysis is presented in the supplementary materials on GitHub.

An interesting correlation between a categorical variable and a numeric variable is the one between the grade and the interest rate (not surprising). Note the outliers in the interest rate (6%). Cohen et al used data reported at different times and found that these interest rates where changed after the loan was issued. Since there are only a few such points we decided to keep them. To address to scale issue, we applied data normalization to our data using the means and the standard deviations from the train data.

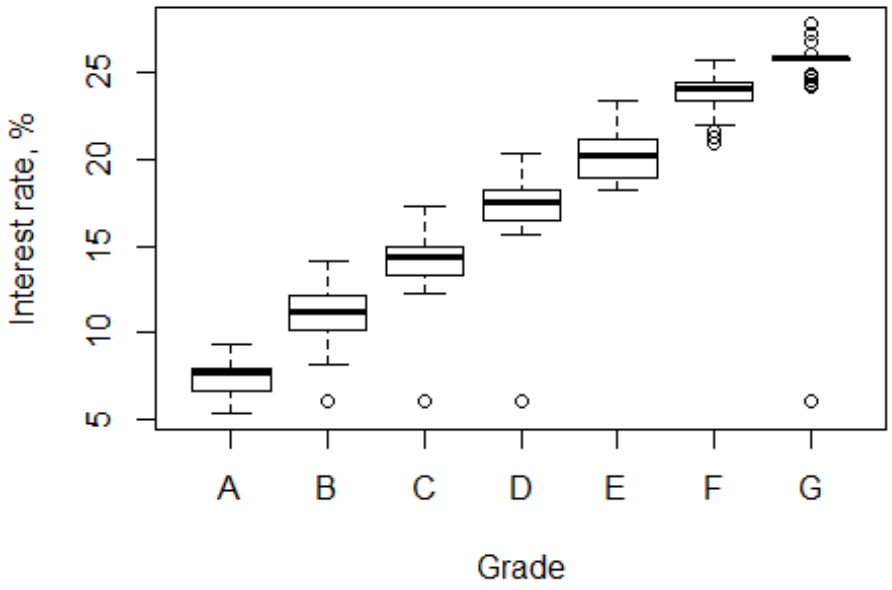

Figure 6. Interest rates assigned to 36-month loans as a function of loan grade

We note a few peculiarities of the data:

- statistical tests indicate strong correlations between the loan status and both numeric and categorical variables;
- conditional histograms and barplots fail to indicate such correlations for most of the variables;
- χ^2 test of independence of the categorical features indicate that all variables are dependent.

To address these issues we used the feature selection using the Boruta package in R and found that: “No attributes deemed unimportant”.

Step 4: Modeling

The focus of this step is to predict defaulting loans using binomial machine learning classification techniques. A typical workflow includes the following steps:

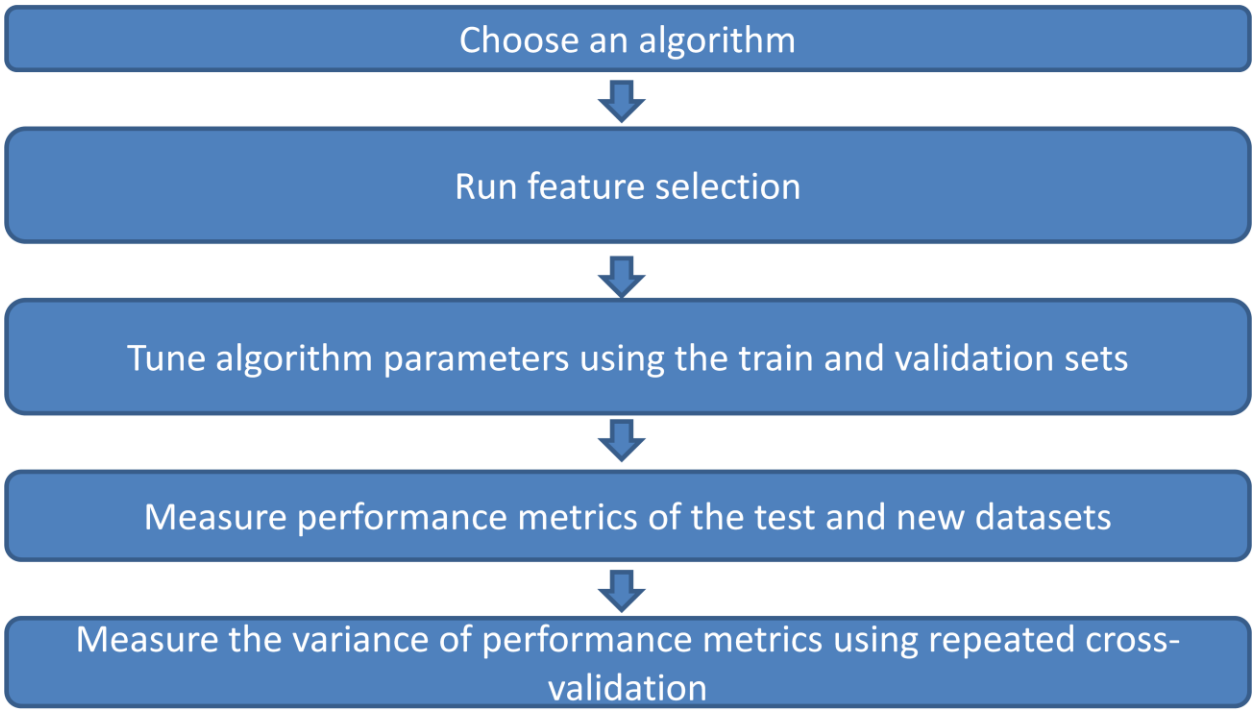

Figure 7. Workflow for binomial loan status classification

For this part of the study we used more than 10 algorithms: Naïve-Bayes, logistic regression, regularized logistic regression (glmnet), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Multivariate adaptive regression splines (MARS), decision trees (rpart), random forests (RF), boosted trees (xgbtree), and artificial neural networks (ANN). We, also, used support vector machines but dropped it because of poor performance.

Most of this work was done in R, but also used Azure ML to test and run comparison between models. For feature selection and parameter tuning we used the functions in the Caret functions in R (for optimization we used AUC). For the logistic regression model used stepwise variable selection. The ANN was performed using the h2o package in R.

The performance of each model on the test and new datasets is presented in Table 3.

Table 3. Results of different classification models on the test and new datasets*

	Naive Bayes		**Logistic Regression**		**RF**		**ANN**		**rpart**		**MARS**		**xgbtree**		**LDA**		**QDA**	
	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New	Test	New
Accuracy	0.82	0.82	0.85	0.84	0.85	0.84	0.85	0.85	0.85	0.84	0.85	0.84	0.85	0.85	0.85	0.84	0.72	0.71
Sensitivity	0.15	0.14	0.00	0.00	0.00	0.00	0	0	0.02	0.03	0.00	0.00	0	0	0.01	0.02	0.36	0.41
Specificity	0.93	0.94	1	1	1	1	1	1	0.99	0.99	1	1	1	1	1	1	0.78	0.76
Precision	0.29	0.33	0.48	0.51	0.44	0.42	NA	NA	0.33	0.35	0.39	0.48	NA	NA	0.43	0.47	0.22	0.24
F1	0.2	0.20	0.00	0.02	0.01	0.01	NA	NA	0.04	0.06	0.00	0.01	NA	NA	0.02	0.04	0.28	0.30
AUC	0.66	0.69	0.67	0.70	0.66	0.68	0.66	0.7	0.62	0.62	0.67	0.69	0.67	0.70	0.68	0.7	0.63	0.65
Brier score	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.24	0.24

(*) The positive class is considered the loan_status = Charged.Off (default)

Brier scores measure how calculated class probabilities are from the actual values. Smaller values mean good indicators.

A quick glance at the data in Table 3, indicate that all models perform poorly at detecting the defaulted loans. When a threshold value of 0.5 is used to assigned labels to class probabilities some models fail to detect a single default customer.

This situation is best illustrated in the ROC curve, as shown below for the test dataset and the glmnet model (alpha = 0.2928571, lambda = 0.007123733).

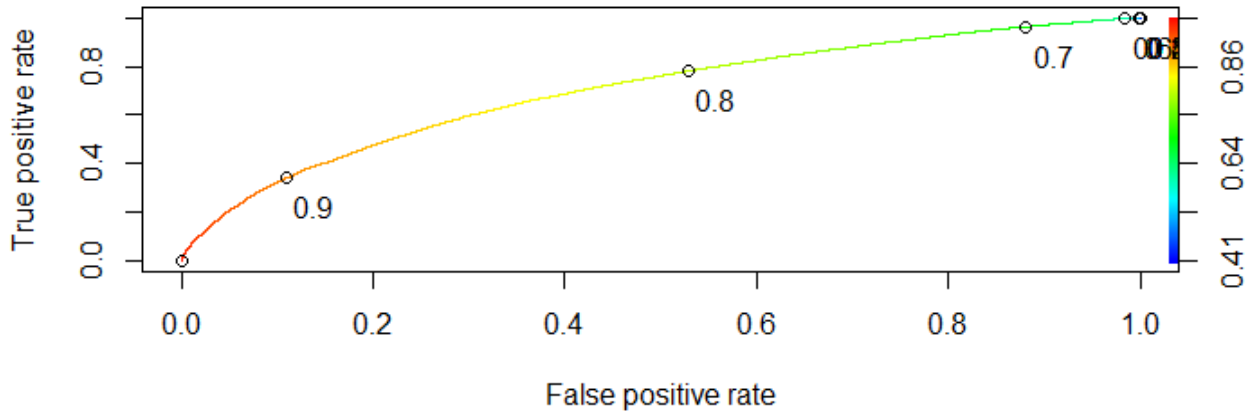

Figure 8. ROC curve for the glmnet model

Here we see that at threshold below 0.6, both tpr and fpr are nearly 1. At a threshold of 0.5, there only 8 true negatives, 4 false negatives, and an accuracy of 0.85. One can increase the true negative rate (default cases) by increasing the threshold value. As shown in the table below, by increasing the threshold value to 0.9 we a pay a price of less than 5 misidentified fully paid cases for one correctly identified charged off case. Similar result were obtained when we balanced the data using random undersampling or SMOTE. The results were similar to the literature results.(Namvar, et al., 2018)

Table 4. Charged Off cases identification as a function of threshold for the glmnet model

Threshold	True negatives	False negatives	Accuracy	Penalty
0.5	8	4	.853	-
0.6	98	132	.853	1.42
0.7	889	1760	.842	1.98
0.8	4597	13149	0.745	2.86
0.9	9996	44650	0.416	4.47
1.0	11635	67678	0	Inf

This approach can lead to a viable investment strategy.

It is worth noting that the all the models perform similarly of the validation, test, and new datasets. The best AUC values are around 0.68, in agreement with literature values. (Cohen, et al., 2018)

The last step outlined in Figure 7, can also be included in the step 5 (Model Evaluation) of our data pipeline. We run repeated 10-fold cross-validation (10 times) in order to get an estimate of the variance in the metrics used to assess the models. In the end we used the AUC values.

The results are shown in Figure 9.

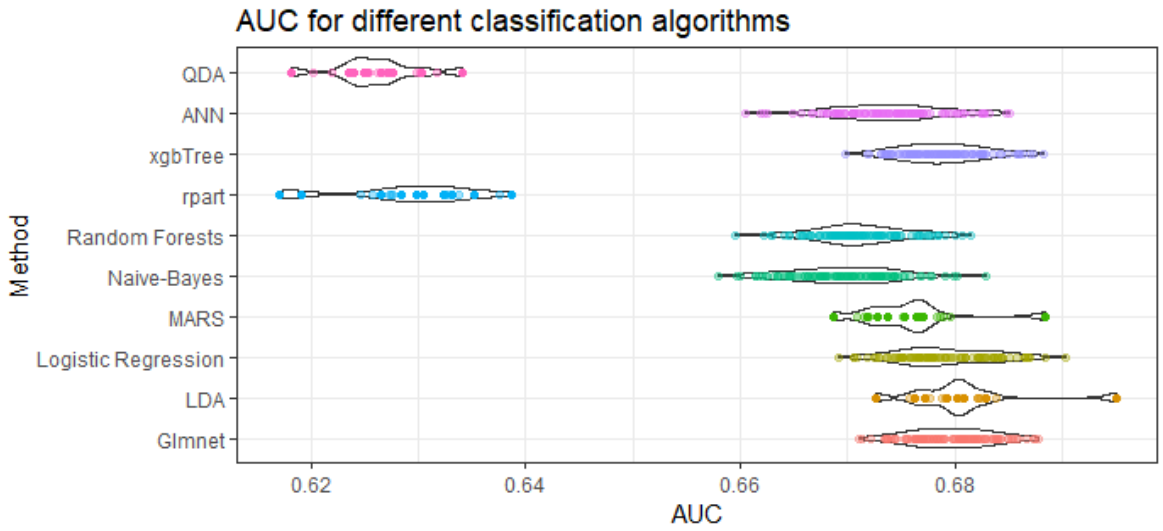

Figure 9. Results of AUC calculation for the binomial models

Using Anova and TuckeyHSD test we conclude that glmnet, logistic regression, xgbtree, and LDA models have the best performance in terms of AUC, with no statistical difference between them. Given that the logistic regression is the only model that retains the feature interpretability, we chose this model to carry on the data analysis.

We made some attempts at using trivariate classification, i.e. classifying the loans into charged_off, prepaid, or fully_paid. Most of the models suffer from data imbalance, as most of the cases belong to the prepaid case. In the end, we choose the glmnet logistic regression model to include in our final hybrid regression – classification model of the ROI.

As hinted in the last paragraph we run linear regression models to fit the ROI for the entire train set, the train set conditioned on the loan_status with two classes: charged_off and fully_paid, and the train set conditioned on the loan status with three classes: charged_off, prepaid, and fully_paid. In the next section we will present the details of the final model.

Step 5: Evaluation

In this step we will examine machine learning based models for the return on investments. The simplest model is a regression model. To add some complexity we added models that combined the probability of default with ROI predicted values conditioned on the loan status value. A schematic diagram of the model is shown in Figure 10.

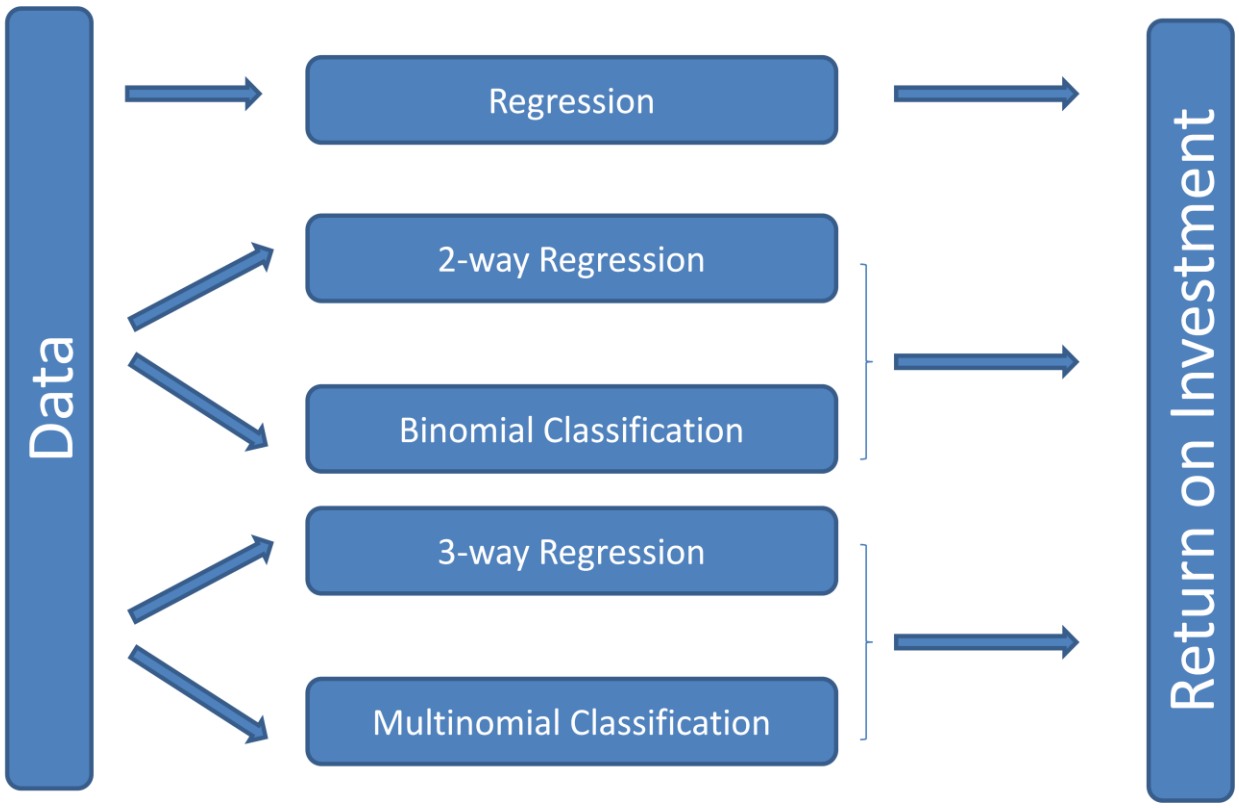

Figure 10. Diagram for return of investment calculation

$$ROI_i = P_i(loan_{status} = charged\_off) * roi_i(loan_{status} = charged\_off) + P_i(loan_{status} = fully\_paid) * roi_i(loan_{status} = fully\_paid)$$

A similar variable is defined for the case of three-class model.

To test the efficacy of the model we designed an experiment where we sampled repeatedly 1000 loans from the test and new data sets (without replacement). From each subsample we picked 100 random loans and calculated the average value of the ROI (random method). For the linear, binary and ternary models we sorted the loans by the ROI and picked the top 100 loans. To complete the comparison we also picked the top 100 loans by the actual ROI (Best choice model, albeit only a hypothetical model). The results are displayed in Figure 11.

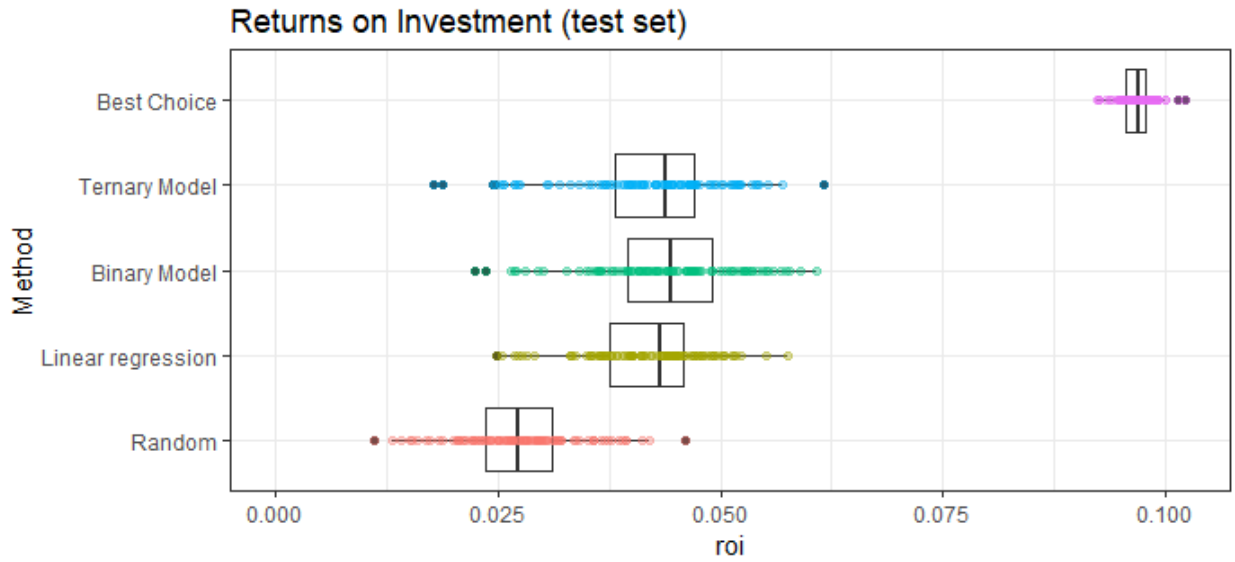

Figure 11. The ROI values for different investment simulations

After running Anova and TuckeyHSD tests we concluded that our models performed significantly better (~ 50 %) compared to the random model, but also significantly worse compared to the model with best possible choice of investments. The actual ROI values are underestimated by this model. We, also, built alternative models where any monthly installment amount received from a borrower is reinvested for the remainder of the 36-month period with the same interest rate or a rate comparable to the Certificate of Deposit rates issued by banks. The model performance was similar to that of the first model.

If one chooses to remain true to the parsimony principle (Ockham's razor), then the linear regression model gives the best performance. It has the advantage of being very portable (just need to store an intercept and some slopes) and it is highly interpretable. The model that uses binary classification is probably more robust and less susceptible to errors.

Step 6: Deployment

One of the key points of the initial proposal was the development of a Shiny application that would connect to Lending Club through an API, pull-up the data, rate each loan request, and buy notes. Unfortunately, at this time, investing requires a Lending Club account (not available for non-US residents).

As an alternative, we deployed a simplified model using a Microsoft Azure free tier account. The user can download the data in an Excel spreadsheet and run the calculation from there.

Conclusions

- used statistical and graphical methods using R and Azure ML to gain insights on the loans offered by Lending Club;
- used a wide range of machine learning algorithms to identify loans likely to default;
- designed and tested investment models that yield better annualized returns on investment;
- created a deployable model using Azure ML.

References

Berger, Sven C., & Gleisner, Fabian. (2009). Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. *Business Research, 2*(1), 39-65. doi: 10.1007/bf03343528

Chapman, Pete, Clinton, Julian, Kerber, Randy, Khabaza, Thomas, Reinartz, Thomas, Shearer, Colin, & Wirth, Rudiger. (2000). CRISP-DM 1.0. Step-by-step data mining guide. Chicago, Il.

Cohen, Maxime C., Guetta, C. Daniel, Jiao, Kevin, & Provost, Foster. (2018). Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science. *Big Data, 6*(3), 191-213. doi: 10.1089/big.2018.0092

Emekter, Riza, Tu, Yanbin, Jirasakuldech, Benjamas, & Lu, Min. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics, 47*(1), 54-70. doi: 10.1080/00036846.2014.962222

Jagtiani, Julapa;, & Lemieux, Catharine. (2018). The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform (pp. 1-32): Federal Reserve Bank of Philadelphia.

Jiang, C. Q., Wang, Z., Wang, R. Y., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. [Article]. *Annals of Operations Research, 266*(1-2), 511-529. doi: 10.1007/s10479-017-2668-z

Larose, D.T., & Larose, C.D. (2015). *Data Mining and Predictive Analytics*: Wiley.

Li, Z. Y., Li, K., Yao, X., & Wen, Q. (2019). Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending. *Emerging Markets Finance and Trade, 55*(1), 118-132. doi: 10.1080/1540496x.2018.1479251

Ma, L., Zhao, X., Zhou, Z. L., & Liu, Y. Y. (2018). A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decision Support Systems, 111*, 60-71. doi: 10.1016/j.dss.2018.05.001

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), 4621-4631. doi: 10.1016/j.eswa.2015.02.001

Martinez-Climent, C., Zorio-Grima, A., & Ribeiro-Soriano, D. (2018). Financial return crowdfunding: literature review and bibliometric analysis. *International Entrepreneurship and Management Journal, 14*(3), 527-553. doi: 10.1007/s11365-018-0511-x

Namvar, Anahita;, Siami, Mohammad;, Rabhi, Fethi;, & Naderpour, Mohsen. (2018). Credit risk prediction in an imbalanced social lending environment. *arXiv:1805.00801*

Polena, Michal, & Regner, Tobias. (2018). Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Class. *Games, 9*(4), 82.

Serrano-Cinca, Carlos, & Gutiérrez-Nieto, Begoña. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113-122. doi: https://doi.org/10.1016/j.dss.2016.06.014

Xia, Yufei, Liu, Chuanzhe, & Liu, Nana. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications, 24*, 30-49. doi: https://doi.org/10.1016/j.elerap.2017.06.004

Zhao, H. K., Ge, Y., Liu, Q., Wang, G. F., Chen, E. H., & Zhang, H. F. (2017). P2P Lending Survey: Platforms, Recent Advances and Prospects. *Acm Transactions on Intelligent Systems and Technology, 8*(6). doi: 72.10.1145/3078848