P2P Loan Data Analysis and Recommendations

Abin Abraham

1 Analysis Domain

Lending Club is the world's largest peer-to-peer lending platform and online credit marketplace enabling borrowers to apply for unsecured personal loans from 1,000 to 40,000 dollars. Investors can browse borrower loan listings and select loans that they want to invest in, based on the information supplied about the borrower, amount of loan, loan grade assigned by Lending Club and loan purpose. Investors make money from interest on the loans. Lending Club earns revenue by billing borrowers an origination fee and investors a service fee.

Investors provide loans with the hope to make a profit from the interest payment by the borrower. There are many different factors that affect a borrower's credit, thus assessing a borrower's credit risk a highly complex task. The underlying motivation for this coursework is determining findings of relevance to model credit risk.

The data is sourced from Kaggle. This anonymized dataset contains data regarding loans borrower's demographics, employment, credit history, loan repayment history and other categorization by Lending Club.

2 Questions

This paper explores the following research questions:

1. What are the factors which influence the status of a loan? Which features explain it the most?
2. To what extent can an investor predict the status of a loan?
3. What are the factors which influence the interest rate for a loan? Which are the most important ones?
4. Can investors identify loans ignoring the grades and still get good returns?
5. What are loans being used for and does it impact the loan defaults?

By pursuing these research questions, we uncover patterns to help Lending Club to help their customers - investors to make better use of the platform for investment and borrowers to understand what drives their loan approval and understand the drivers of the interest rate assigned to their loans.

3 Plan

1. Import the dataset from Kaggle.
2. Data wrangling
    A. Original dataset has more than 2 million records majority of which are current loans for which the outcome may change in the future and thus may not be suitable for analysis.
    B. Handling missing values and imputation.
    C. Cleaning of values in annual income and employment length
    D. Dropping of features with posterior knowledge of our analysis target – loan status.
3. Data transformation
    A. Log transformation
    B. One- hot encoding
    C. Standardization
4. Exploratory data analysis and insights
    A. Univariate and multi variate data analysis
    B. Assess collinearity and remove highly correlated columns
5. Feature Engineering
6. Train and build classification models to understand the underlying phenomenon and assess feature importance
7. Assess models and compare results

4 Analysis

The information available for each loan consists of all the details of the loans at the time of their issuance as well as more information relative to the latest status of loan. The data has features which are nominal, ordinal, ratio and interval data-types.

Almost 40 % of the features in the dataset has more than 80% null values and thus was not be suitable for data analysis. In addition, there are multiple

columns which will only be available, after a loan is processed. We have dropped them. For rest of the features, imputation using medians have been completed.

Loan Status Analysis

Firstly analysing the loan status distribution, around 78 % of the data is for full paid loans and thus good loans. Rest of them can be considered potentially bad loans.

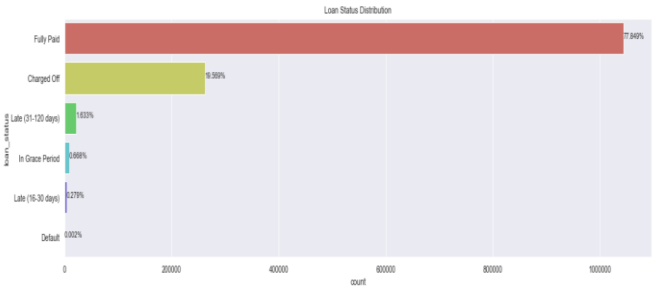

Figure 1: Loan Status Count Distribution

The loan amount distribution is multi-modal with two biggest peaks at 10K and 15K. The distribution of loan amount is slightly skewed to the right. Based on the box plot we can suggest that loan amounts for bad loans are generally a bit higher than good loans.

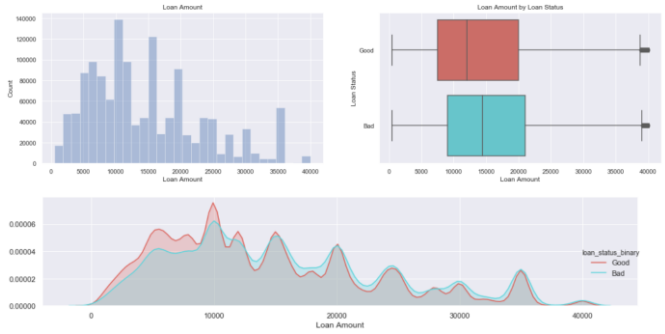

Figure 2: Loan Amount histogram, box plot and kernel density plot

Moving on to interest rates, the distribution of interest rate is also slightly skewed to the right. We have an interesting pattern here. Higher interest rates seem to be a good indicator of bad loans. We have many outliers in the interest rate > 25% for good loans and above 27% for bad loans, but this data may be important to understand the credit risk behavior. Investors who are willing to take the risk are rewarded high interest rates. Skewed features are log transformed.

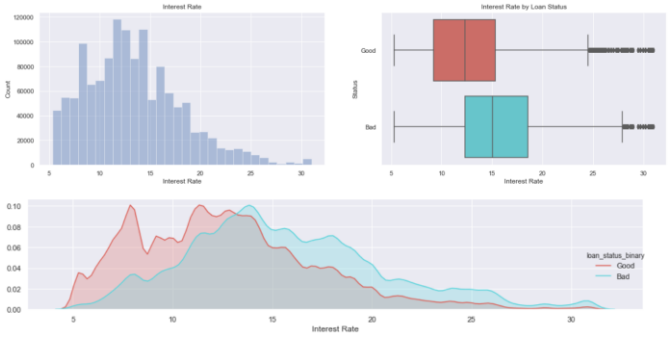

Figure 3: Interest rate histogram, box plot and kernel density plot

Analysing trends behind the loan purpose we notice that debt consolidation and credit card loans have an overwhelming majority of purpose in loans. Loans with purpose of small_business, renewable_energy and moving have the highest bad loan rates while wedding, car and home improvement have the least.

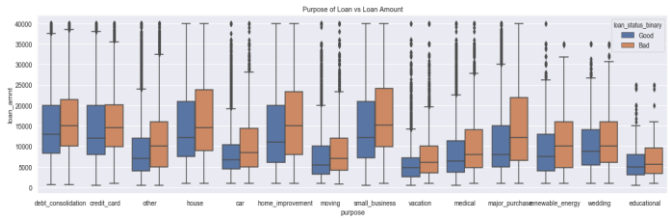

Figure 4 Loan amount - Purpose- Status Box Plot

The initial Loan grade calculated by Lending club https://www.lendingclub.com/foliofn/rateDetail.action accurately reflects the expected loan status of the loan in the future. The proportion of bad loans increase almost linearly as per the initially assessed loan grade i.e. proportion of bad loans for Grade A < B< C < D < E < F. This also opens our eyes to the unexpected bad loans for the high graded loans. Joint applicants have higher proportion of bad loans.

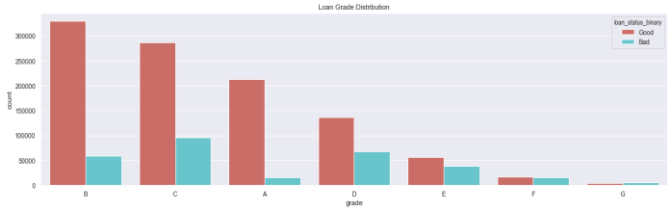

Figure 5 Loan Purpose - Status bar-chart

Considering the outliers identified while exploratory data analysis, we have used mahalanobis distance and removed rows with mahalonabis distance greater than critical value. Next, we create new

features for loan amount to annual income ratio and credit line age.

Since there are close to 30 features let us switch to understand the collinearity through a correlation plot

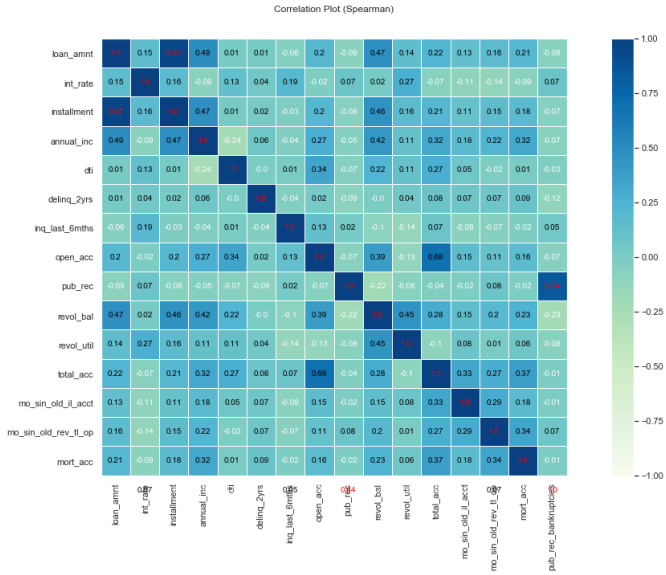

Figure 6 Correlation Chart

We have high collinearity between-loan_amnt and instalment & total_acc and open_acc & credit_history_age and mo_sin_old_rev_tl_op. We drop the latter values in each case.

Based on modelling using various algorithms we have poor goodness of fit / F1 score suggesting difficulty in modelling, but the key drivers excluding loan grades are: log_int_rate, log_loan_amnt,employmentlength, purpose_ small_business and home_ownership_RENT. Thus investors can rely on other loan attributes.

Interest Rate Analysis

Regression Analysis can be used both for the purpose of prediction or inferential statistics. In our case, we use it s to identify the impact of features on interest rate:

Params 2:

OLS Regression Results

Dep. Variable:	log_int_rate	R-squared (uncentered):	0.271
Model:	OLS	Adj. R-squared (uncentered):	0.271
Method:	Least Squares	F-statistic:	1.097e+04
Date:	Tue, 10 Dec 2019	Prob (F-statistic):	0.00
Time:	05:27:43	Log-Likelihood:	-1.2684e+06
No. Observations:	1005729	AIC:	2.537e+06
Df Residuals:	1005695	BIC:	2.537e+06
Df Model:	34		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.953e-15	0.001	-2.29e-12	1.000	-0.002	0.002
emp_length	0.0191	0.001	21.572	0.000	0.017	0.021
dti	0.0754	0.001	79.779	0.000	0.074	0.077
delinq_2yrs	0.0801	0.001	92.625	0.000	0.078	0.082
inq_last_6mths	0.2378	0.001	274.127	0.000	0.236	0.240
pub_rec	0.0442	0.001	39.433	0.000	0.042	0.046
revol_bal	-0.0436	0.001	-44.778	0.000	-0.046	-0.042
revol_util	0.3220	0.001	349.782	0.000	0.320	0.324
total_acc	0.0126	0.001	12.117	0.000	0.011	0.015
mo_sin_old_il_acct	-0.0484	0.001	-51.207	0.000	-0.050	-0.047
mort_acc	-0.0266	0.001	-25.182	0.000	-0.029	-0.025
pub_rec_bankruptcies	0.0595	0.001	52.691	0.000	0.057	0.062
credit_history_age	-0.1062	0.001	-111.087	0.000	-0.108	-0.104
loan_income_ratio	-0.0099	0.001	-11.649	0.000	-0.012	-0.008
log_loan_amnt	0.2324	0.001	227.041	0.000	0.230	0.234
log_annual_inc	-0.1976	0.001	-173.860	0.000	-0.200	-0.195
home_ownership_ANY	0.0032	0.001	3.746	0.000	0.002	0.005
home_ownership_MORTGAGE	-0.0244	0.001	-43.172	0.000	-0.026	-0.023
home_ownership_NONE	0.0020	0.001	2.316	0.021	0.000	0.004
home_ownership_OTHER	-0.0006	0.001	-0.721	0.471	-0.002	0.001
home_ownership_OWN	0.0091	0.001	12.704	0.000	0.008	0.011
home_ownership_RENT	0.0191	0.001	33.000	0.000	0.018	0.020
purpose_car	-0.0010	0.001	-1.216	0.224	-0.003	0.001
purpose_credit_card	-0.1479	0.001	-233.602	0.000	-0.149	-0.147
purpose_debt_consolidation	0.0145	0.001	27.724	0.000	0.013	0.016
purpose_educational	-0.0017	0.001	-1.990	0.047	-0.003	-2.57e-05
purpose_home_improvement	0.0246	0.001	31.138	0.000	0.023	0.026
purpose_house	0.0371	0.001	43.809	0.000	0.035	0.039
purpose_major_purchase	0.0163	0.001	19.677	0.000	0.015	0.018
purpose_medical	0.0416	0.001	49.581	0.000	0.040	0.043
purpose_moving	0.0465	0.001	54.864	0.000	0.045	0.048
purpose_other	0.1029	0.001	130.093	0.000	0.101	0.104
purpose_renewable_energy	0.0139	0.001	16.312	0.000	0.012	0.016
purpose_small_business	0.0645	0.001	76.992	0.000	0.063	0.066
purpose_vacation	0.0336	0.001	39.532	0.000	0.032	0.035
purpose_wedding	0.0110	0.001	12.900	0.000	0.009	0.013

Omnibus:	1558.866	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1826.528
Skew:	-0.037	Prob(JB):	0.00
Kurtosis:	3.195	Cond. No.	2.28e+15

Figure 7 OLS Regression Analysis with interest rate as dependent

We have a low R2 of 0.271 which suggest we have relatively poor goodness of fit, but the model explains the variability around the mean for the interest rate.

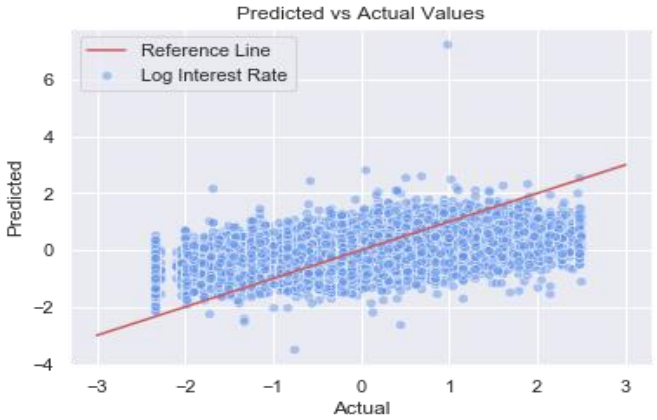

Figure 8 Predicted v/s Actual

The key drivers- Number of inquiries since last 6 months, revolving account utilization, log loan amount, purpose as other increases the interest rate while purpose of credit card, log annual income and credit history age reduces it.

Modelling and Dimension Reduction
We have performed a 3-fold cross validation on a sample from the dataset.

Model	Accuracy	F1
LogisticRegression	0.78	0.12
GaussianNaiveBayes	0.28	0.36
KNNeighbors	0.75	0.23
DecisionTree	0.68	0.31
RandomForest	0.77	0.17

Based on the results we selected LogisticRegression and GaussianNaiveBayes and attempted to train using 75% of the dataset and test using 25%.

Model	Accuracy	F1
LogisticRegression	0.78	0.14
GaussianNaiveBayes	0.72	0.40

Overall the results for the F1 score have been poor suggesting that it is challenging to predict the loan status accurately. Post feature selection F1 has improved to 0.41 for GNB.

There were around 50 data variables and we used PCA to reduce dimensionality while preserving data variability. There was poor explained variance of 6% in each of the first 2 components and no clear separability between classes.

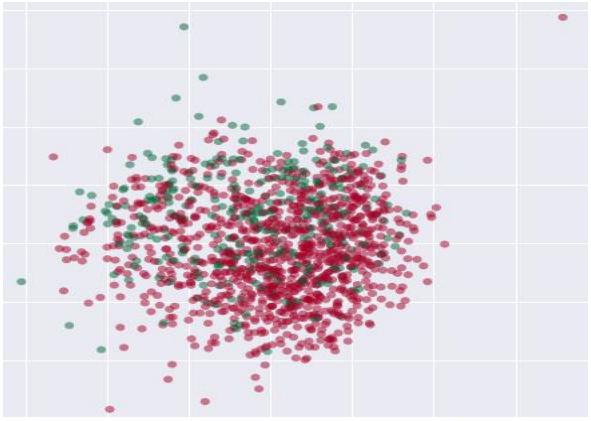

Figure 9 PCA scatter plot first 2 components

5 CRITICAL REFLECTION

Based on our study, Lending Club can vision a tool for borrowers to evaluate their loan profile, before making a formal request for a loan to understand their chances of having it granted. A formal application for a loan usually affects the user's credit score. The intelligence gained though this analysis can be used to build this new tool benefitting them without affecting their credit score.

Our study can also benefit investors. Currently most rely on the loan grade alone to decide to invest. Investment asset selection is both art and science. With our study we have laid the foundation for loan investment selection without relying on loan grades, but this needs more data to substantiate and finally compare with the loan grades

We cannot always reliably model a phenomenon using available data. Usually, occurrences that are a proxy for human behavior are difficult to predict compared to natural processes or system behavior like interest rate. Modeling loan status is tantamount to understanding human behavior as people can default on a loan despite having the means to pay for it and people may still pay in full and on time despite having high debt to income ratio and a history of delinquencies. Probably there isn't a pattern. More data attributes to analyze may help to understand the behavior. Alternatively, segmenting the population first based on certain parameters can be used to model their repayment behavior. We could observe the goodness of fit for interest rate which is system generated.