Lending Club – Analyzing the drivers for interest rate

Data Science Program, SICE, Indiana University

Atharva Tere, Nilima Sahoo, Piyush Shinde, Saheli Saha

Introduction

Lending club is a peer-to-peer lending company which connects investors with customers seeking loans. The project focuses on the analysis of the data and the visualization to understand the trend of customers and various factors that are influencing Lending Club customers. We are trying to find answers of the trend of the loan, variation according to the region, factors driving the interest rates and loan status.

Data Exploration and Insights

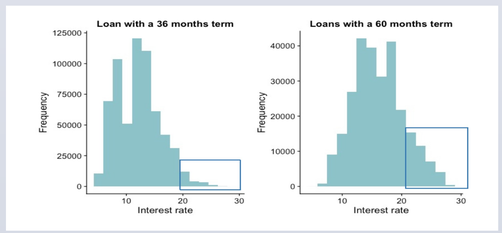
Loan Purpose

Word clouds are not generally considered the best way to visualize text data, but it seems to be a good option in this case because we have keywords instead of text data and hence there is no noise. Also, our emphasis here is on the frequency of word occurrence, which is best highlighted by a word cloud.

We can learn from the word clouds that –

1. The first word cloud explains the purpose for loan application. The majority is credit card bills or debt consolidation.
2. The second word cloud talks about the people who are applying for the loans. It looks like the people at mid-senior levels are applying for loans more prominently.

Loan Term



As we see, higher interest rates are in more proportion in the long term loans as compared to the short term loans. The x-axis is interest rate while the area denotes the number of loan applications. Even if the range of y-axis is different, the distributions still say that higher interest rates are more prominent in the long term loans as compared to short term loans.

Regression model to predict interest rates

We created a linear regression model to gain insights on the features affecting loan interest rate.

The model shows that the borrowers - annual income, (B & C) loan grade, (0 & 1) inquiry in credit file, (credit card & debt consolidation) loan purpose, total current balance and 36 month term loans inversely affect the loan interest rate while rented home ownership, and verified verification status directly affect it.

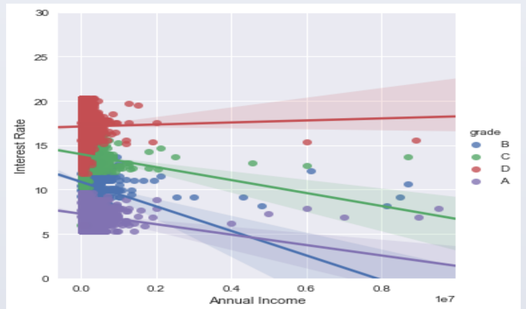
In-Depth analysis of Lending Club Data

Root MSE	3.42594	R-Square	0.3886
Dependent Mean	13.24644	Adj R-Sq	0.3886
Coeff Var	25.86311		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	33.38206	0.08228	405.72	<.0001	0	.	0
ln_annual_inc	1	-1.15074	0.00809	-142.24	<.0001	-0.13874	0.72428	1.38068
grade_B	1	-2.39082	0.00896	-266.72	<.0001	-0.24680	0.80483	1.24250
grade_C	1	-0.41559	0.00886	-46.91	<.0001	-0.04245	0.84150	1.18835
home_ownership_RENT	1	0.45677	0.00865	52.78	<.0001	0.05110	0.73508	1.36039
inq_0_0	1	-2.35050	0.01024	-229.60	<.0001	-0.26622	0.51258	1.95092
inq_1_0	1	-1.15778	0.01135	-102.03	<.0001	-0.11761	0.51865	1.92810
purpose_credit_card	1	-1.87818	0.01156	-162.44	<.0001	-0.18104	0.55474	1.80265
purpose_debt_consolidation	1	-0.58392	0.00993	-58.79	<.0001	-0.06553	0.55463	1.80300
term__36_months	1	-3.75095	0.00836	-448.81	<.0001	-0.39233	0.90178	1.10892
ln_tot_cur_bal	1	-0.22989	0.00400	-57.50	<.0001	-0.06323	0.56984	1.75487
verification_status_Source_Verif	1	0.72262	0.00912	79.26	<.0001	0.07969	0.68162	1.46710
verification_status_Verified	1	1.79072	0.00939	190.76	<.0001	0.19188	0.68110	1.46821

Loan Grades and Annual Income

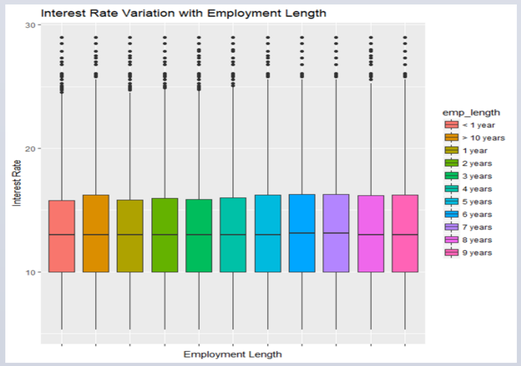


We can clearly see how the interest rate varies with the loan grades (A, B, C & D) pre-defined by the Lending Club.

Interest rates for loan grade D are greater than 15%, while for A it is lesser than 10%.

Unfortunately there was not much variation between annual income and loan interest rates, whereas we gathered a few insights from the borrowers annual income. The median annual income is $67,000 and 75% of the annual income lies in the range of $50,000 to $94,000.

Employment Length

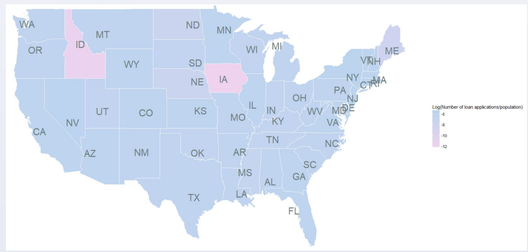


In the plot of loan interest rate vs employment length, we can see there is absolutely no variation of interest rate.

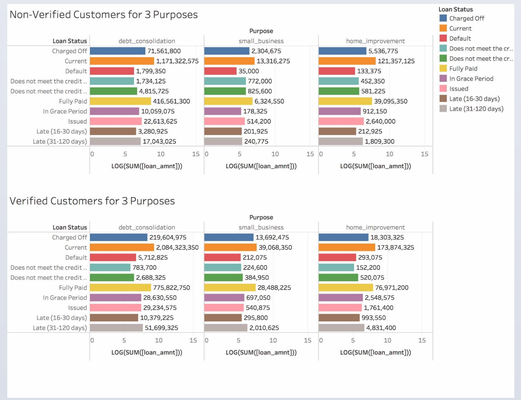
Thus, we can conclude that the loan interest rate is independent of length of employment.

Distribution of loan applications

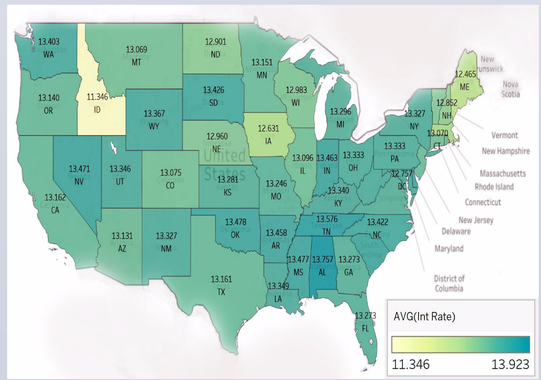
The image below shows the variation of number of loan applications from a state with respect to population in that particular state. It can be inferred from the map that not many people from Idaho and Iowa are applying for loans from Lending Club as compared to other states.



According to the plot below, it is clearly visible that verification status does not make a big difference in Loan Amount. Verified and non-verified customers are getting loans of same amount irrespective of their verification status.

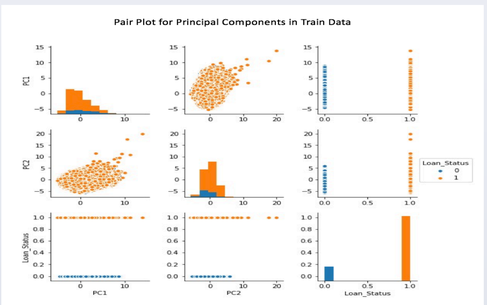


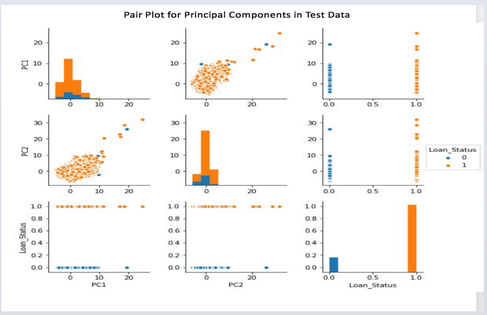
The graph below shows variation of interest rate according to different states. It shows that people in Idaho and Iowa are paying low interest rates. It could be because, there are less people applying for loans from these states. It also shows people in Atlanta are paying higher interest rates.



Conclusion

- From the word cloud visualization we can see maximum customers are of manager level and they have applied for loans mostly with a purpose of debt consolidation and credit card.
- We were focused on analyzing various factors which are effecting interest rate such as term, annual income, employment length.
- Through a regression model to predict the interest rate we found out features affecting it.
- For predicting loan status for future customers we tried to apply classification model and used PCA for feature selection. This did not work because the PCAs extracted were not linearly separable.





References

1. Ian Fellows. *Wordcloud*: Word Clouds, 2014. R package version 2.5
2. Kurt Hornik Ingo Feinerer and David Meyer. *Tm*: text mining, 2008. R package version 0.7-1
3. Hadley Wickham. *Ggplot2*: Elegant Graphics for data analytics, 2009
4. *Plotly* Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2015 URL: https://plot.ly

Acknowledgement

We would like to thank our Professor, Dr. Yong-Yeol Ahn and the team of associate instructors for their continuous support and guidance throughout the semester that helped us build the fondant ions of visualization. We would also like to thank SICE for giving us a chance to showcase our work in the research symposium.