International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1977–1986

Exploratory Visual Anatomy of Prosper Loan Dataset Using Tableau for P2P Lending

Shubhankar Gupta[1], Saksham Gupta[2], Arpan Singh[3], Bindu Garg[4]

[1]Student, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering Pune, Maharashtra, India
[2]Student, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering Pune, Maharashtra, India
[3]Student, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering Pune, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, Bharati Vidyapeeth College of Engineering Pune, Maharashtra, India

Abstract

Prosper is a money lending platform which helps in reducing the distance between the borrower and lender. We have done exploratory analysis on the relationship between the stakeholders that directly affects the borrowers' Prosper Score and who come under the defaulters. This work aims to show the facts about loan data and emphasize on loan borrowers, occupations, annual income and their APR. The Prosper Dataset includes information about the borrower's background and detailed description of the loans. The borrowers' Annual Percentage Rate (APR) will be analyzed on the basis of many factors such as borrower's rating, score, occupation and income that could have a direct impact on his APR. This work also aims to investigate type of borrowers and the type of loans taken by them. We have inspected the dataset and used Tableau to create our visualizations in order to find out the defaulters, reason for defaults and the cause for the borrowers to apply for loan.

Keywords- *Borrower's APR,Defaulter Rate,Prosper Loan Dataset,Prosper Score,Tableau, Exploratory Visual Analysis, Univariate Exploration*

1. Introduction

Numerous investors have made profit using this platform and countless borrowers are able to get more money easily. Even though it provides credit score and basic information of borrowers to keep a check on the safety of the transactions yet one cannot deny the risk of large number of people losing their money. [10]

1.1 The Prosper loan data

The Prosper loan dataset comprises of 113,937 loan entries with 81 attributes on each loan, including loan amount, interest rate, and status of the loan taken and income of the borrower from the year 2009-2014[9]. There are four types of variables

1) **Loan Status**: It consists of the status of the loan which is taken such as Past Due, Final Payment in progress, Completed etc.
2) **Borrower Data**: Basic attributes of the borrowers such as annual income, condition of employment, etc.
3) **Loan Data**: Summary of the loan such as loan tenure and the annual percentage rate (APR)
4) **Credit Risk Metrics**: Metrics evaluating the risk associated with the loans such as Prosper score and bank card utilization.

ISSN: 2233-7857 IJFGCN

	ListingKey	ListingNumber	ListingCreationDate	CreditGrade	Term	LoanStatus	ClosedDate	BorrowerAPR	BorrowerRate	LenderYield
108301	78AB3383182033739820413	104341	2007-02-27 19:36:21.373000000	D	36	Completed	2008-07-15 00:00:00	0.20936	0.2020	0.1920
104425	865434103616570524A7AE5	271052	2008-01-24 19:53:18.283000000	HR	36	Defaulted	2010-06-29 00:00:00	0.31363	0.2900	0.2800
31606	72C03588532410539535B37	902200	2013-09-17 06:11:10.480000000	NaN	36	Past Due (91-120 days)	NaN	0.23898	0.2015	0.1915
2728	2BD43466545609013767851	430809	2009-10-28 19:51:53.480000000	NaN	36	Completed	2010-07-29 00:00:00	0.12967	0.1085	0.0985
14204	D8063479117253337678 2C1	452218	2010-03-30 18:48:21.053000000	NaN	36	Completed	2011-02-17 00:00:00	0.37453	0.3500	0.3400
19206	AAB43532104380330B32FFC	539877	2011-11-18 17:25:22.780000000	NaN	36	Current	NaN	0.18986	0.1609	0.1509
47810	44E93366860755133E657A1	34338	2006-08-26 09:42:25.647000000	A	36	Completed	2007-03-07 00:00:00	0.11244	0.0950	0.0900
28766	BE4A353159180810338E0CB	539397	2011-11-16 13:05:02.210000000	NaN	36	Completed	2012-07-25 00:00:00	0.19088	0.1619	0.1519
96484	2B8F35985124700361B70DB	1066378	2013-12-29 01:25:46.067000000	NaN	60	Current	NaN	0.22549	0.2010	0.1910
63941	B196352583644416661DFE1	528004	2011-09-18 17:03:59.517000000	NaN	36	Completed	2013-07-02 00:00:00	0.35643	0.3199	0.3099

Fig. 1 Prosper dataset

1.2 Components of Prosper Rating

It is calculated by Estimated Loss Rates which is deduced using Prosper Score and Credit score.

a) Prosper Score: It diversifies between 1 and 11, with 11 being the minimum risk and 1 being the maximum risk.

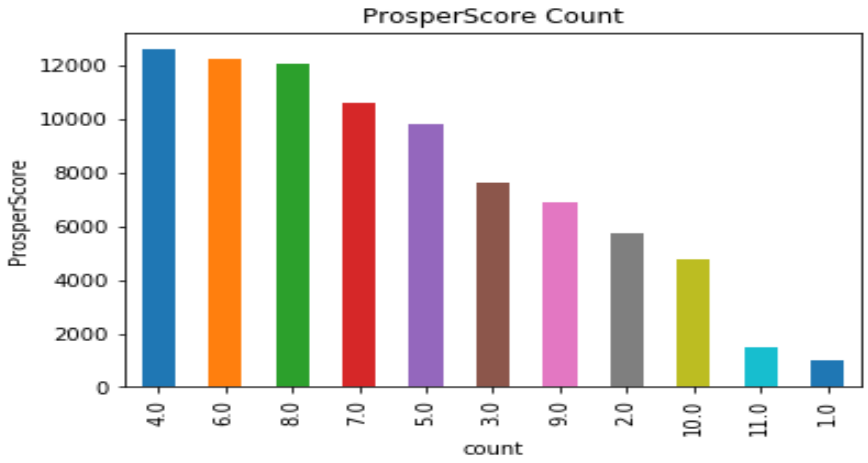

Fig. 2 Prosper score count

b) Credit Score Average: It is computed by taking the mean of CreditScoreRangeLower and CreditScoreRangeUpper, both variables being related to the Prosper Score. [9]

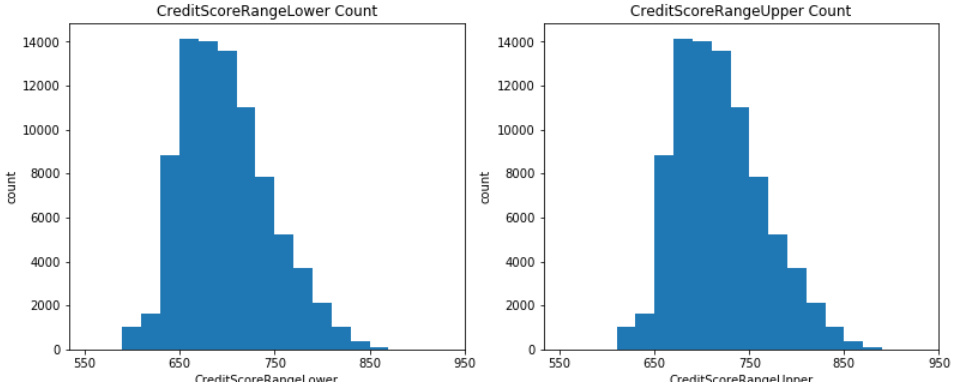

Fig. 3 Credit score average

ISSN: 2233-7857 IJFGCN

2. Literature Review

P2P lending platform is relatively new as it was started in 2005. There are many banking standards that are studied by the researchers as well as many small or big abbreviations or parameters that have to be kept in mind while performing research on it like the geographical and demographic character of the borrower to take into account all feature before making any decision and to reduce the chance of losses.[21]

The previous work in 2011 was done to understand the trends given by the p2p platform as well as to identify the scope of improvement thereafter to understand the adoption curve for such a platform.

The case study adopted for doing research is Paipaidai the largest Chinese p2p platform done in 2013. They used the collected dataset which is provided by the company then they classify based on different parameters like different types of loans like home loan, car loan, business loan, etc. given to different types of borrowers. it gives the raw analysis of the fact that which type of loan is more opted by the borrower. This gives the clients behavior regarding the specific type of loan.[11]

In India, the recent years there is an increase in the number of people applying for loans. So, now in this particular paper, exploratory visual analysis is a backbone as in this research paper based on overall predictions is done based on different types of graphs like a bar graph, pie chart, etc.

The method proposed by them is simple but yet efficient as we inculcate the input which would act as the raw data then it will move towards the completion of data then raw data would be refined so that unwanted data could be removed then after obtaining the refined data the date would be classified and at last, the final data which they obtained from the classification would be updated and the cycle continues.[15]

Using this method the loan sanctioning mechanism is eased as well improved now based on the other parameter the graph is created after studying the graph it was found that through result which they obtain after analyzing graph were inconsistent and having some defects because of missing values so remove this abnormality they applied some algorithm in it like KNN as it helps in classifying the dataset based on the similarities in the dataset and to make data more consistent and to remove the abnormalizes binning algorithms are used. another algorithm is used that is naïve byes algorithm and the combination of these 3 would give the scope or chance for more improvised prediction.[19]

Through the researches done in 2014 and 2015 with the more improvised technique, it was found that shorty term loan is preferred by the borrowers.

They also found that the data are more accurate when the above factors are combined with tableau as it plays a more intuitive role in the finding the characteristics of the borrower as well as provides the confidence among the lenders so more sophisticated but more practical banking operations could be obtained by various above researches. But every research has scope for improvement so in the upcoming researches we will see a more robust method with the help of data analytics, statistical and visual approach with tableau and python Libraries.

3. Overview of the proposed model

Taking into account the attributes of the dataset, the system uses exploratory analysis to find patterns among dataset and find relationship among variables and the strength of variable in determining some of features such as borrower’s APR. This analysis is done in various iterative steps from collecting raw data to finding pattern from preprocessed data.

ISSN: 2233-7857 IJFGCN

A. Data Pre-Processing-

This includes converting raw data to structured data that can be used for analysis and finding patterns it includes following components-

Data cleaning-It is process of refining data and removing noisy and irrelevant data from the dataset such as removing null entry, cleaning missing values, detecting and removing data discrepancy. It is done to increase the accuracy of the model. [5]

Feature selection-It is the procedure for extracting relevant data for analysis and only the attributes/variable needed for analysis is selected and all the other attributes are removed.

Data transformation-It is procedure of modifying data into suitable form required by data analysis and mining algorithms. It is generally done using data mapping and code generation techniques. In our work we have converted numerical (dimension) to categorical (measures) and vice-versa wherever required.

B. Data Analysis-

It is process of understanding the data, structure of data, number of entities and data types of variables and then doing analysis on data to find out interesting patterns from data, grouping data.

Data aggregation-In this process data is gathered and expressed in summary form like climbing up the concept hierarchy for representing information in brief or expressing more details for statistical analysis. Our analysis also aggregates data based on mathematical functions like sum, median and also on time-series analysis like monthly. It also helps to gather specific information about some particular groups based on any variable like income. [6]

Univariate and Bivariate Exploration-Univariate analysis is one of the most basic form of analysis and deals with only one variable and do not take into account any relationship or causes. It summarizes data and find patterns in data.

It is quantitative analysis of two variables for finding out the empirical relationship between the two variables. It can be descriptive or inferential analysis. Our work uses correlation matrix to find the strength of relationship among variable. [8]

C. Pattern Recognition-It involves extracting/mining pattern from the analysis done in above steps. It also involves transformation of task related data to patterns and finding out cause-effect diagram. It also uses some algorithms like decision tree, strength induction, characterization, etc. It is followed by pattern evaluation finding its importance using p-score or other evaluation techniques. [2]

D. Knowledge Representation-It is process of summarization and visualization to make user understand data. It involves use of different data visualization tools like in our analysis we have used tableau to display the findings, generate tables, reports. In tableau for our analysis we have used scatter plots, line plots, and geo-maps, bar plots, state-wise maps, and making an interactive dashboard for better visualization. [7]

ISSN: 2233-7857 IJFGCN

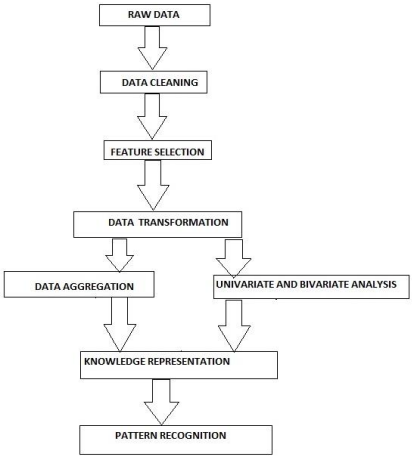

Fig. 4 Proposed model

4. Prosper Dataset Analysis

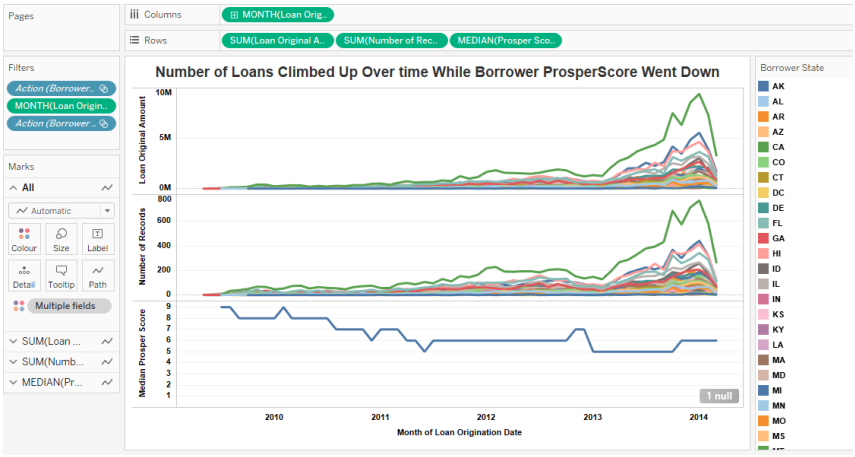

Fig. 5 Number of loans vs prosper score

In the above line plot we can clearly see the different borrower state from USA while the plot is judged among three parameters namely:- loan original amount , no of records and no of prosper score respectively as the state the defaulters rate is progressively increasing due to which money lender are unwilling to invest in the end which results in decline of prosper score.

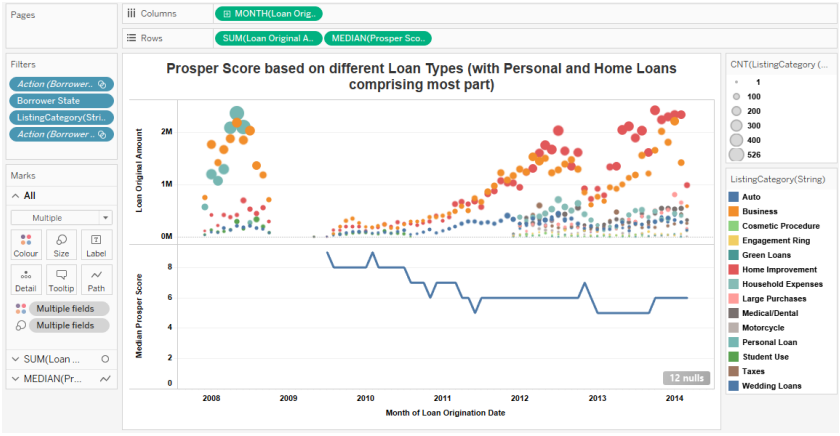

Fig. 6 Relationship between prosper score and loans

ISSN: 2233-7857 IJFGCN

While in the above scatter plot within a span of four years the loan amount reached 2m and most of the people opt for home improvement then followed by business and so on.

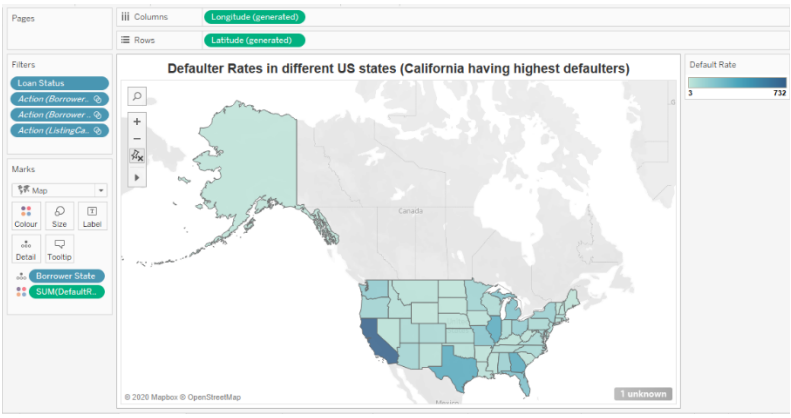

Fig. 7 Defaulter rates in different U.S.A. states

In the above geo map of USA, it is clearly visible that most of the defaulter arrives from California so the prosper score relatively for that state is minimized.

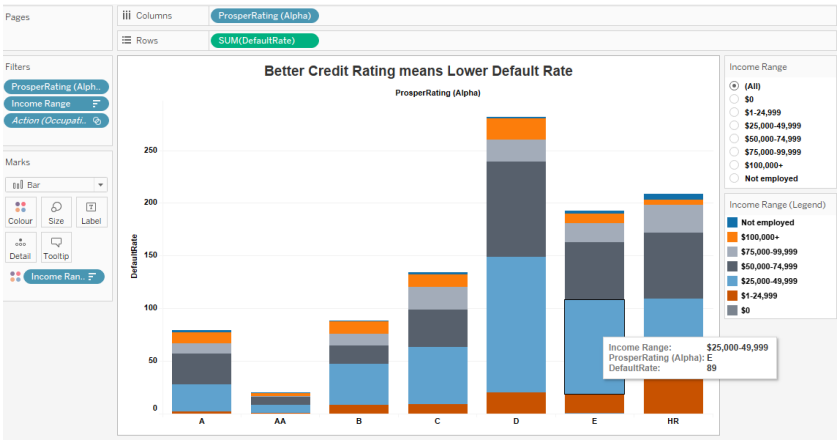

Fig. 8 Credit rating vs default rate

In the above bar graph, it could be understood that if credit rating is high for an individual then his chances of becoming defaulter is reduced as he will repay the loan in a mentioned tenure.

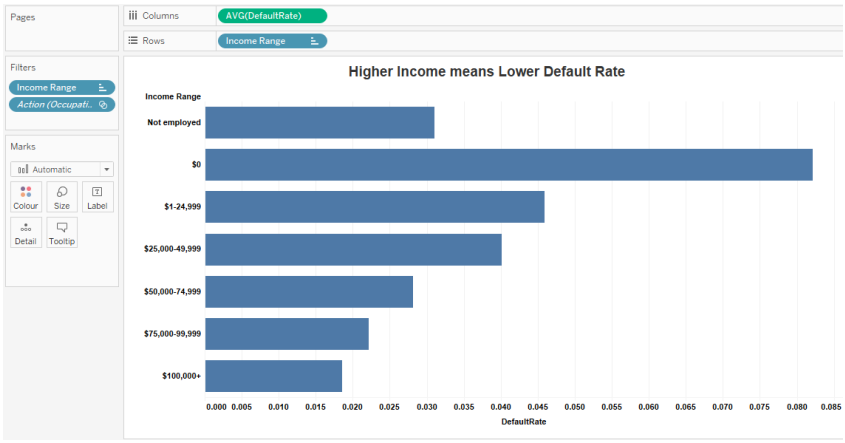

Fig. 9 Income vs Default Rate

ISSN: 2233-7857 IJFGCN

The above graph clears our concept regarding the relationship between higher income and default rate. For an individual if the income is high so the risk of becoming defaulter is reduced as the chances of repayment of dues increased the major risk is for the unemployed individual.

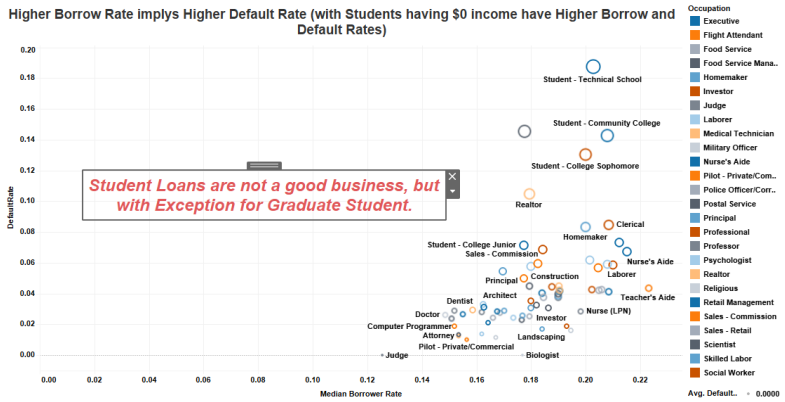

Fig. 10 Relationship between borrow rate and default rate

The above plot mentions about the demerits of student loans as it increases the chances of non-repayment of loans as it totally depends on the condition that if the student remains unemployed so there are the chances that an individual is not in the condition to repay his loan.

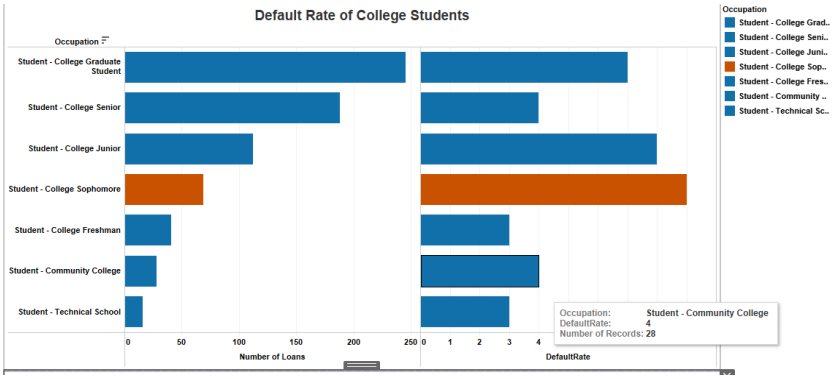

Fig. 11 Default rate of college students

The above graph hints about the possibility of if the individual is in higher grades then the chances of becoming employed are more as compared to the students in lower grades so if they are employed then the risk of becoming loan defaulter is reduced.

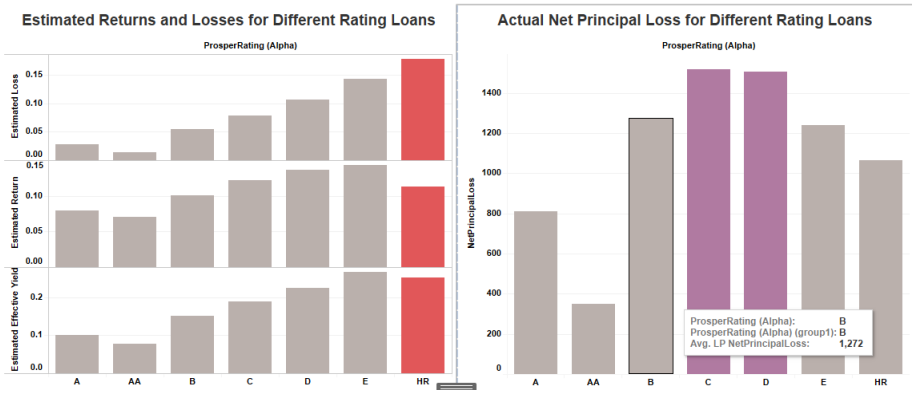

Fig. 12 C&D loans

ISSN: 2233-7857 IJFGCN

The graph indicates about the risk involved in investing in C&D loans as the chances of repayment of the loans are minimal due to which there is net principal loss but at the same times higher instantaneous return could be expected.

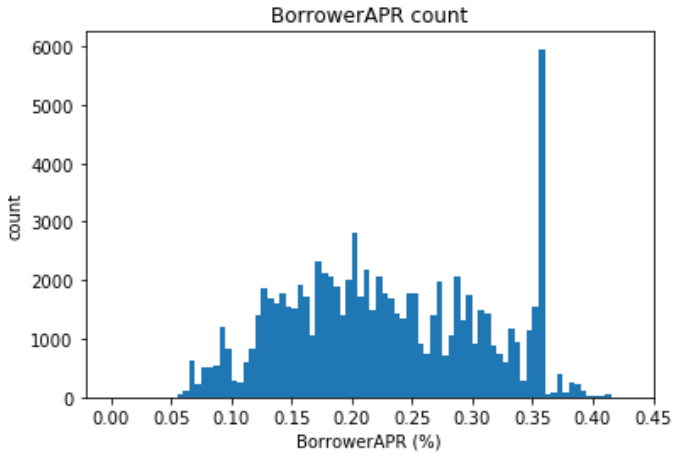

Fig. 13 Borrower's APR count

Borrower APR is termed as the annual percentage rate charged from borrower in annual basis .it could be in the form of any kind fee etc. so in the finance sector the borrower is privilege with the comparison of rates provided by different lenders.

Fig. 14 Correlation plot

The table mentions about the relationship among the different parameters or analyzing the relations among the values for example if we see the first column that is borrower APR and compare it with credit score then it is inversely proportional.

5. Conclusion

The time series analysis from 2007 to 2014 depicts the number of loans taken by the borrowers, the amount of their respective loans and the way in which their Prosper Score varied during this tenure. We can clearly see that since 2009, the transactions escalated since 2013 and then plummeted at

ISSN: 2233-7857 IJFGCN

beginning of 2014. The borrower credit scores showed a constant drop rate during this interval and some states had their default rates more than 30%.

We found that the students have the highest default rates and most of them invested in the loan type 'D'. Considering their occupations, we were able to find out a unique pattern that the college students that signed up for higher studies have more loans to their name as well as showed the highest default rates. Also, the sophomore (juvenile) students were negligent and have borrowed less amount of loans. Considering the return on different categories of loans, the loan graded in 'HR' incurred the extreme losses irrespective of the credit score of the borrowers. However, loans 'C' and 'D' are the riskiest ones.

An interesting pattern was discovered that California has comparatively higher number of defaulters (>700) as compared to Texas, New York and Illinois.

In Bivariate analysis Out of all the attributes, the variable Prosper Score showed negative correlation with respect to Borrower APR. The Credit Score Range Lower, Available Bank Card Credit and Credit Score Range Upper are all positively correlated to Prosper Score and negatively correlated to Borrower APR.

References

[1] Battle, L., & Heer, J. (2019). Characterizing Exploratory Visual Analysis:A Literature Review and Evaluation of Analytic Provenance in Tableau. Eurographics Conference on Visualization (EuroVis), 15.

[2] Jency, X. F., Sumathi, V. P., & Sri, J. S. (Novembor 2018). An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients. International Journal of Recent Technology and Engineering (IJRTE), 4.

[3] Kacheria, A., Shivakumar, N., Sawkar, S., & Gupta, A. (september 2016). Loan Sanctioning Prediction System. International Journal of Soft Computing and Engineering (IJSCE), 4.

[4] Medvedev, S. (2019, jun 12). Exploratory Data Analysis with Tableau.

[5] Yen, L. (2019, jan 4). P2P Lending Platform Data Analysis: Exploratory Data Analysis.

[6] Zhanga, Y., Li, H., Hai , M., Li, J., & Li, A. (2017). Findings of loan funded successful in P2P Lending system. ELSEIVER, 6.

[7] Zhao, Hongke & Ge, Yong & Liu, Qi & Wang, Guifeng & Chen, Enhong & Zhang, Hefu. (2017). P2P Lending Survey: Platforms, Recent Advances and Prospects. ACM Transactions on Intelligent Systems and Technology. 8. 1-28. 10.1145/3078848.

[8] Barasinska, N. (2009). The role of gender in lending business: Evidence from an online market for peer-to-peer lending, The New York Times (2009:217266), pp. 1-25. Berger, S.C., and Gleisner, F. (2007). Electronic marketplaces and intermediation: An empirical investigation of an online p2p lending marketplace.

[9] Garman, S., Hampshire, R., and Krishnan, R. (2008a). Person-to-person lending: The pursuit of competitive credit markets, Proceedings of the International Conference on Information Systems, Paris, pp. 1-16.

[10] Herzenstein, M., Andrews, R.L., Dholakia, U., and Lyandres, E. (2008). The democratization of personal consumer loans? Determinants of success in onlinepeer-to-peer lending communities.

ISSN: 2233-7857 IJFGCN

[11] Legris, P., Ingham, J., and Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model, Information &Management (40:3), pp. 191-204.

[12] Leong, Carmen, Barney Tan, Xiao Xiao, Felix Ter Chian Tan, and Yuan Sun. (2017) "Nurturing A Fintech Ecosystem: The Case of a Youth Microloan Startup in China." Int J Inf Manage 37: 92–97.

[13] Einav, L., Jenkins, M., and Levin, J., 2013. The impact of credit scoring on consumer lending. RAND Journal of Economics , 44(2): 249–274.

[14] Zhang, Shaofeng, Wei Xiong, Wancheng Ni, and Xin Li. (2015) "Value of Big Data to Finance: Observations on an Internet Credit Service Company in China." Financ Innov 1: 17.

[15] Zhang, Chenghong, Tian Lu, and Tian Lu. (2017) "Assessment of Borrowers ' Delinquency and Default Behaviors in Online P2P Lending : A Two- stage Model", Twenty First Pacific Asia Conf. Inf. Syst.

[16] Ding, H., P. Zhang, T. Lu, H. Gu, and N. Gu. (2017) "Credit Scoring Using Ensemble Classification Based on Variable Weighting Clustering", in 2017 IEEE 21st Int. Conf. Comput. Support. Coop. Work Des. pp. 509–514.

[17] Xia, Yufei, Xiaoli Yang, and Yeying Zhang. (2018) "A Rejection Inference Technique Based On Contrastive Pessimistic Likelihood Estimation for P2P Lending." Electron Commer Res Appl 30: 111–124.

[18] Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 2015, 47, 54–70

[19] Bachmann, Alexander & Becker, Alexander & Buerckner, Daniel & Hilker, Michel & Kock, Frank & Lehmann, Mark & Tiburtius, Phillip & Funk, Burkhardt. (2011). Online Peer-to-Peer Lending – A Literature Review. Journal of Internet Banking and Commerce. 16.

[20] Lin, M., Prabhala, N. R., & Viswanathan, S. (2009b). Judging borrowers by the company they keep: social networks and adverse selection in online peer-to-peer lending. papers.ssrn.com. College Park.

[21] Livingston, L., & Glassman, T. (2009). Creating a new type of student managed fund using peer-to-peer loans. Business Education & Accreditation, 1(1), 1-14.

[22] Klein, T.(2008). Performance in Online Lending Platforms. Online. Friedrich-SchillerUniversität Jena.

[23] Mo S, Chen KC, Ye C. The Evolving Role of Peer-to-Peer Lending: A New Financing Alternative. J Int Acad Case Stud. 2016 May 1;22(3).

ISSN: 2233-7857 IJFGCN