Lending Club Case Study

Executive PG Programme in Machine Learning & AI - January 2023

Name: Bharat Hegde
Email: bharathhegde2005@gmail.com

26th Feb 2023

Introduction

The problem is related to a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

The problem statement is to come up with the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment. The driving factors should be such that they can be utilised during the process loan grant decision itself.

Overall Approach

The company has shared a data file which contains the performance data of the loans in past many years. As part of this case study:

- Understand the structure of the data
- Clean the data
- Analyse and Visualise the data
- Conclusion: Suggest driving factors behind the loan default to the company

The python language is used for all above steps. The libraries used as *numpy*, *panda*, *matplotlib* and *seaborn*

Understand the structure of the data

- A "loan'csv" (loan data) and a "Data_Dictionary.xlsx" (description of columns in loan data) are provided.
- The loan.csv has loan data for past 39717 loans, and each loan has 111 features
- Following is the divide of borrowers who Fully paid in the past vs Defaulters (Charge d Off) vs Currently paying.
    - `Fully Paid     32950`
    - `Charged Off     5627`
    - `Current         1140`

Cleaning the data

- **Removal of unnecessary columns:** The columns which contained only 'NA' or 0 or only one value in the entire column. After pruning the column, the number of feature columns left are: 46
- **Converting strings to Ints :** The interest rate and revolving balance contained % in the value which needed to be removed, also converted to int. Also the term contained additional string "months" which needed to be removed.
- **Replacing null values with 0:** The null values in *mths_since_last_delinq* and *emp_length* were filled with zero.
- **Formatting the values for consistency:** The additional strings such as "years", "year", "+", "<" were removed in the *emp_length* column and values converted to ints.
- **Convert to numerical category codes for Correlation Analysis:** Additional columns contained numerical values mapping to following features were created for Correlation analysis*: verification_status, purpose, sub_grade, home_ownership, loan_status, grade, addr_state*

Analysing and Visualisation of the data

Correlation Matrix and Driving Factor shortlisting

The Correlation matrix is created, and we are interested in how various facotrs correlate to the Loan Status.

Following factors seemed to reasonably correlate with the loan status and we will investigate them further.

`grade_code (0.211936) –` LC assigned loan grade

`term(0.211644) –` The number of payments on the loan. Values are in months and can be either 36 or 60.

`int_rate(0.226822) –` Interest Rate on the loan

`inq_last_6mths(0.067612) –` The number of inquiries in past 6 months (excluding auto and mortgage inquiries)

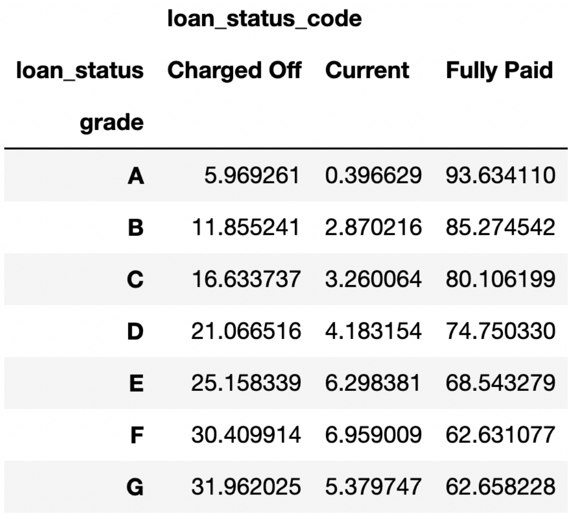
`purpose(0.040771) –` A category provided by the borrower for the loan request.

`revol_util (-0.10)` Also found that, during the tenure if the Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit, increases that seems to have some correlation with loan status

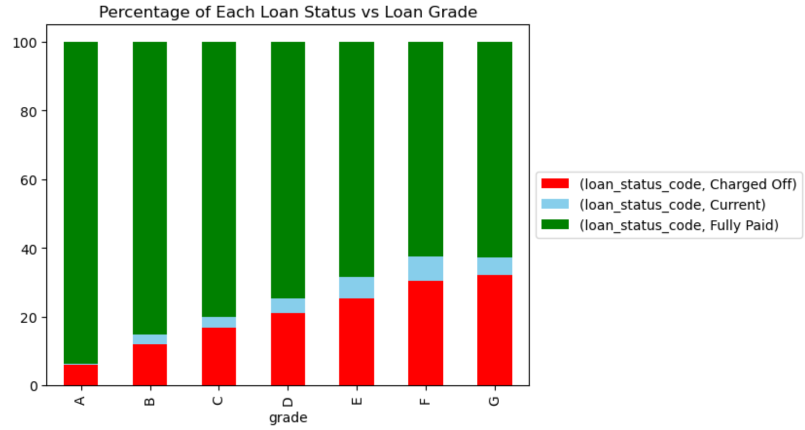
Also, Initially had suspected that *emp_length, dti, delinq_2yrs* could be significant factors as well, but they turned out to be insignificantly related to the eventual loan status

Impact of Loan Grade

(Values are in percentage)



	loan_status_code		
loan_status	**Charged Off**	**Current**	**Fully Paid**
grade			
A	5.969261	0.396629	93.634110
B	11.855241	2.870216	85.274542
C	16.633737	3.260064	80.106199
D	21.066516	4.183154	74.750330
E	25.158339	6.298381	68.543279
F	30.409914	6.959009	62.631077
G	31.962025	5.379747	62.658228

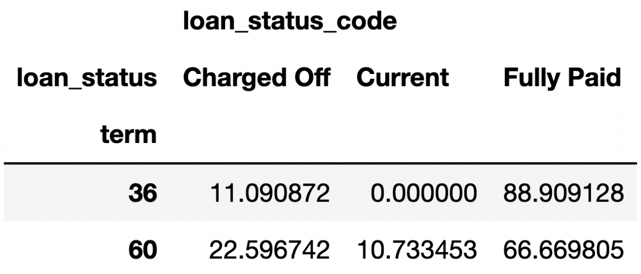


Conclusion 1:

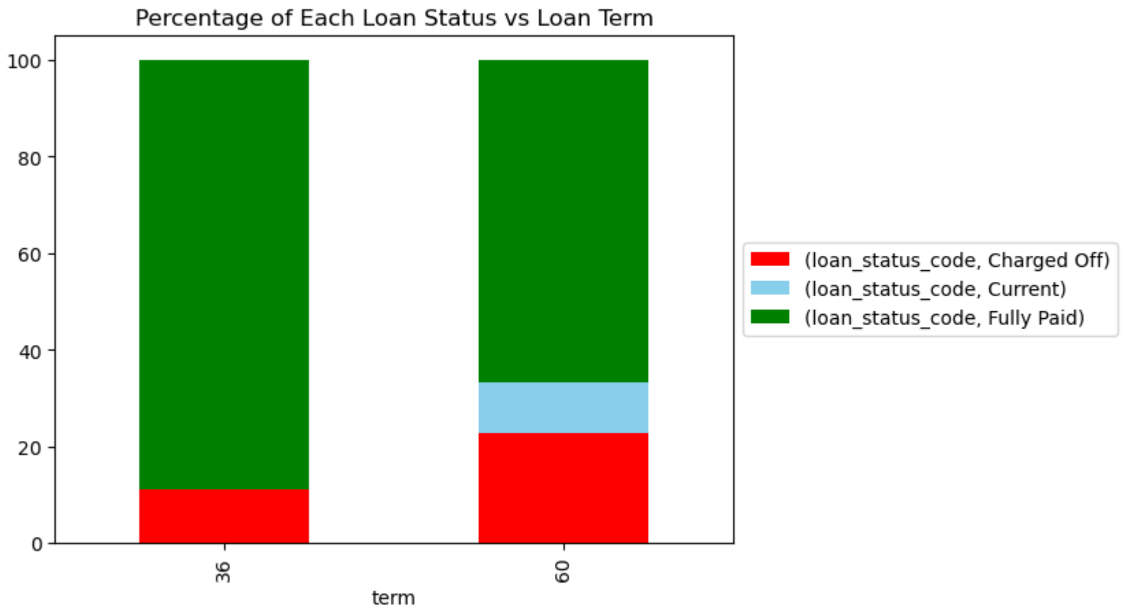
Grade A (6% defaulters) and B (11% defaulters) relatively safer. Other grades: C (16.6% defaulters), D(21.0% defaulters), E(25.1% defaulters), F(30% defaulters), G(31.96% defaulters) are risky, since default % is comparatively higher.

Impact of Loan term

(Values are in percentage)

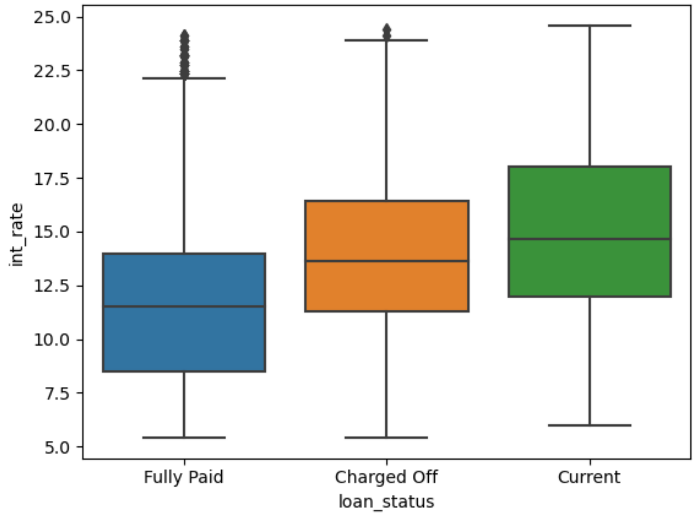


	loan_status_code		
loan_status	**Charged Off**	**Current**	**Fully Paid**
term			
36	11.090872	0.000000	88.909128
60	22.596742	10.733453	66.669805



Conclusion 2 :The loan Term = 36 months is much safer (11% defaulters) compared to term = 60 months (22.6% defaulters)

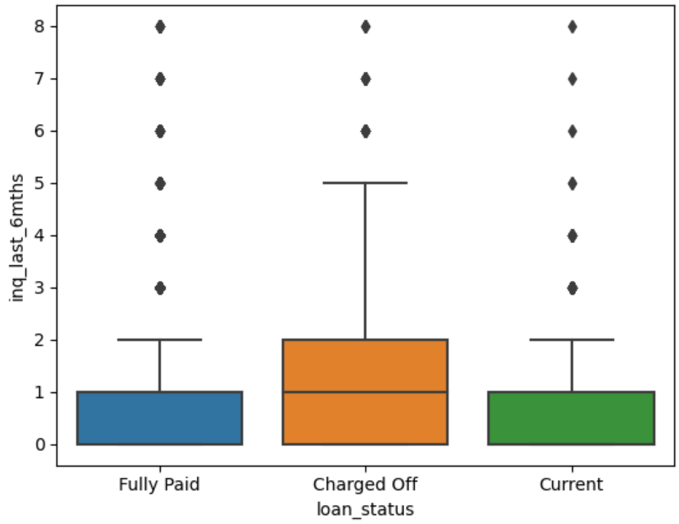
Impact of Interest Rate



Conclusion 3 : As interest rate increases, probability of borrowers repaying decreases. For pully paid the mean int_rate ~= 12% and for defaulters mean int_rate is roughly 13.8%(Ignoring the currently paying members, since their population is lower, and we don't know if they were defaulters later)

25 percentile, 50 percentile value, and 75 percentile value for Default cases is more than that of Fully Paid.

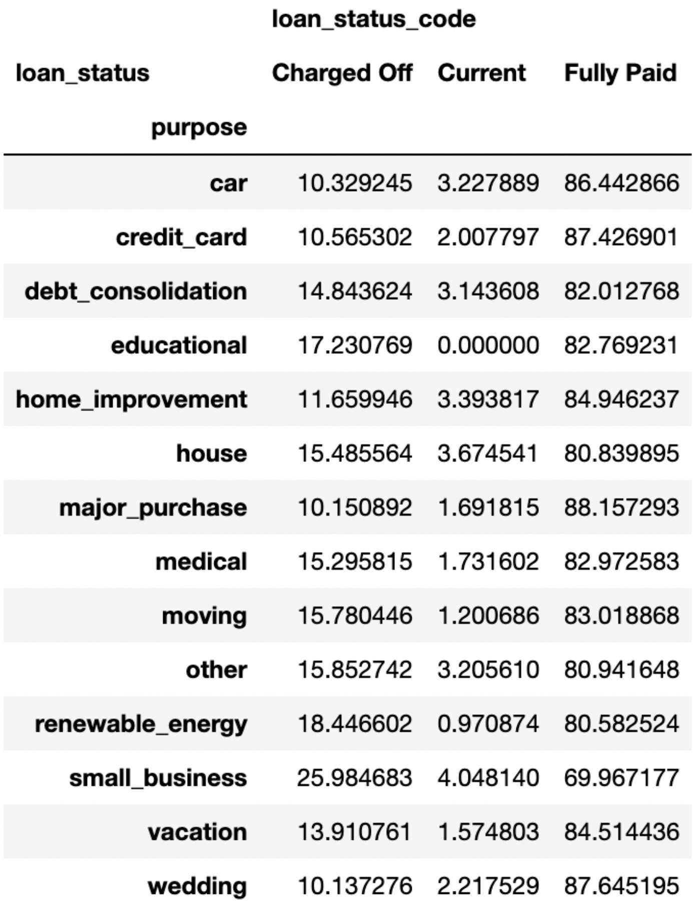
Impact of the Number of inquiries in last 6 months



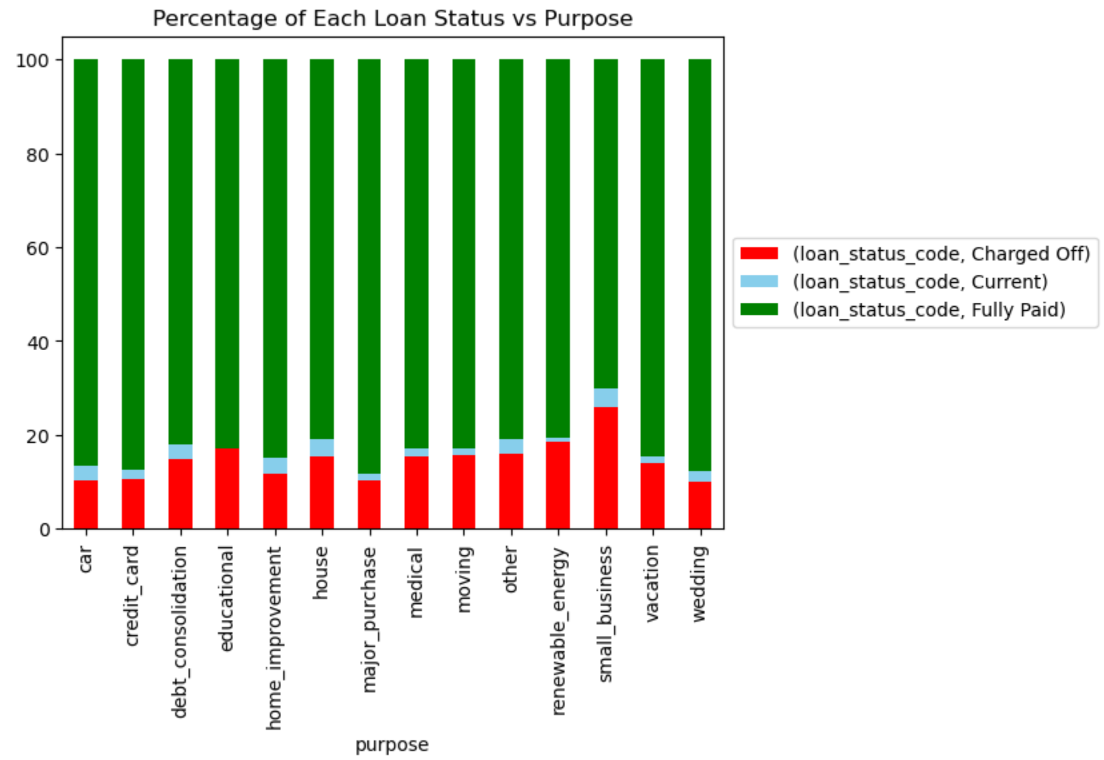
Conclusion 4: As number of inquiries in last 6 months increases, the chances of default also decreases. As per the above data, if number of inquiries is more than 1, then chances of repaying decreases.

50 percentile value, and 75 percentile value for Default cases is more than that of Fully Paid.

Impact of the Purpose mentioned

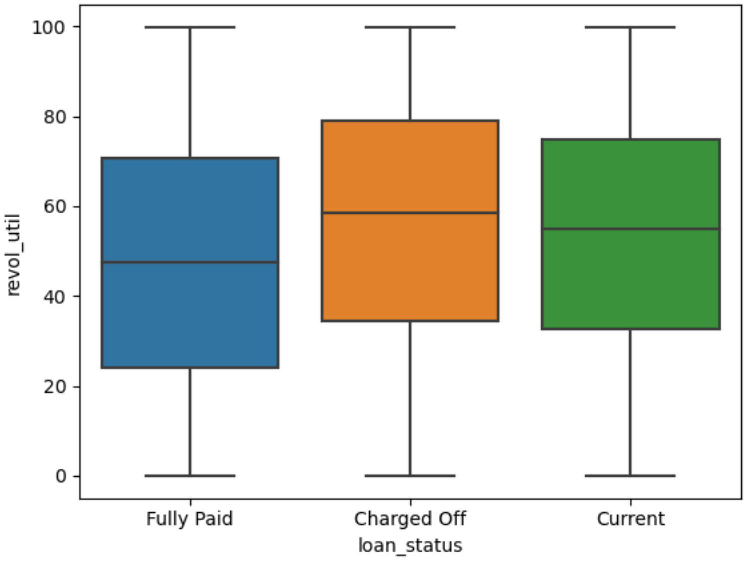


	loan_status_code		
loan_status	Charged Off	Current	Fully Paid
purpose			
car	10.329245	3.227889	86.442866
credit_card	10.565302	2.007797	87.426901
debt_consolidation	14.843624	3.143608	82.012768
educational	17.230769	0.000000	82.769231
home_improvement	11.659946	3.393817	84.946237
house	15.485564	3.674541	80.839895
major_purchase	10.150892	1.691815	88.157293
medical	15.295815	1.731602	82.972583
moving	15.780446	1.200686	83.018868
other	15.852742	3.205610	80.941648
renewable_energy	18.446602	0.970874	80.582524
small_business	25.984683	4.048140	69.967177
vacation	13.910761	1.574803	84.514436
wedding	10.137276	2.217529	87.645195



Conclusion 5 : If somebody mentions purpose of loan as Small Business, then there is 26% chance of defaulting. Other high risk categories are: Educational, House, Medical, Renewable Energy (with probability ~= 15%)

Impact of the Revolving Utilisation Balance



Conclusion 6 : (While the loan is in force) As Revolving Utilisation of the bank balance increases, the chances of default also increases.

As per the above data, if revol_util is more than 50%, then chances of repaying decreases.

25 percentile, 50 percentile value, and 75 percentile value for Default cases is more than that of Fully Paid.

Summarizing Recommendations to the company

1. Grade A (6% defaulters) and B (11% defaulters) are relatively safer. Other grades: C (16.6% defaulters), D(21.0% defaulters), E(25.1% defaulters), F(30% defaulters), G(31.96% defaulters) are risky, since the defaulting % is comparatively higher. **– This is the major deciding factor**

2. The loan Term = 36 months is much safer (11% defaulters) compared to term = 60 months (22.6% defaulters)

3. As interest rate increases (above 13.8%)probability of borrowers repaying decreases.

4. As number of inquiries in last 6 months increases, the chances of default also decreases. As per the data, if number of inquiries is more than 1, then chances of repaying decreases.

5. If somebody mentions purpose of loan as Small Business, then there is 26% chance of defaulting. Other high risk categories are: Educational, House, Medical, Renewable Energy (with probability ~= 15%)

6. Also during the tenure of the loan**,** As the Revolving Utilisation of the bank balance increases by more than 50%, the chances of default also increases. When company gets indications around, company needs to take suitable measures to avoid the default.