Lending Club Case Study

Team Members

1. Wajdi Tahmoush
2. Fadi Anjrou

Lending Club –Case Study

Table of Content

- Problem Description
- Analysis Approach
- Data understanding
    - Overall Loan Status
    - Loans situation understanding
- Data Cleaning and Manipulation
- Data Analysis:
    - Defaults
        - Defaults By Interest Rate
        - Defaults By loan Purpose
        - Defaults By Borrowers income
        - Defaults By ratio of amount to income
        - Defaults By Revolving line Util Rate
        - Defaults By prior bad record
        - Defaults By debts to income
- Recommendations

THE PROBLEM DESCRIPTION

Company Information

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.
- Borrowers can easily access lower interest rate loans through a fast online interface.

Company Objectives from EDA

- the company wants to understand the driving factors **(or driver variables)** behind loan default, <u>i.e. the variables which are strong indicators of default.</u>
- The company can utilize this knowledge for its portfolio and risk assessment.

The company seeks to find solid basis to accept or reject applications in order to increase business and reduce losses from rejecting low risky applications or accepting high risky ones

Problem Statement & Analysis Approach

Problem Statement

Accepting or rejecting loan applications carries 2 types of financial risks:

1- Rejecting loans for applicants who are likely to repay results in a loss of business to the Club.

2- Approving the loan applicants who are likely to default (not repaying) will result in a financial loss for the Club.

The study aims to analyze applicants' probability of repaying or defaulting, and provide the Club with a solid basis to accept or reject applications, and decid on the proper interest rates.

Study Methodology The study is based on a lending history dataset provided by the Lending Club. The study analyzes the data set and provide methods for measuring the default probability of applicants based on different criteria. The study is completed in 3 major steps as follows:

1- Analyze the data set and decide on the variables that can serve in the analysis study; remove all useless variables to reduce ambiguity and focus only on useful data.

2- Clean/fix data values in important variables and generate new matrices if needed in preparation for the analysis.

3- Analyze the different variables and their relation to default probability.

Understanding Data Set

Import the Data Set

Import the data set and check the content. Noticed 39,117 rows and111 columns of Data

Import the Data Dictionary

Revise the provided data dictionary to understand content

Segmenting the Data (Loans Status)

Noticed that provided data contains 3 major categories of loans:

1. Current Loans: These are loans that are still active, and can’t be considered as Defaulted or Not.
2. Fully Paid: These are closed loans that are fully paid (Good Loans).
3. Charged Off: These are the Default Loans (bad Loans) that cause losses for the company.

As our analysis aims to predict defaults, Active loans can’t be considered in the study and should be removed.

The data set has a big portion of columns with no data and should be cleaned.

Data Cleansing and Manipulation

Different Steps has been applied on data in preparation for analysis

1. Remove all Active Loans.
2. Check for duplicate records; No duplicates found.
3. Drop all columns with no data; (Columns reduced from 111 to 57 Variable)
4. Identify relevant and irrelevant variables to the study; Variables related to the current loans or customer behavior like payments, delinquent, etc… are irrelevant to the study. Variables to define the client, or the loan characteristics are those to consider.
5. With the help of the data dictionary file and the provided variables descriptions and values, we decided on relevant variables created a variables file with Relevant flag.
6. Dropped all irrelevant variables. (Variables reduced from 57 to 28.
7. Check for variables not important for the study and drop them like Id, Member Id, etc.…
8. Final Data set has 38577 rows and 21 columns.
9. Understand the columns and their content: Categorical vs Continuous variables.
10. Create New columns flagging Good and Bad Loans based on the Loan Status.

Data Analysis

Having all the data cleaned and ready for Analysis

Analyze Default against Categorical Variables

1. Created a generic function for analyzing Categorical variables and their relation to Good/Bad Loans.
2. The relation is defined as the percentage of good and bad loans in each Category.
3. Run the Analysis function against each Categorical variable and visualize the results using Pie, Bar charts.

Analyze Default against Continuous Variables

1. Created a generic function for analyzing Continuous variables and their relation to Good/Bad Loans.
2. The relation is defined as the percentage of good and bad loans in each Category.
3. Run the Analysis function against each Continious variable and visualize the results using line Charts.

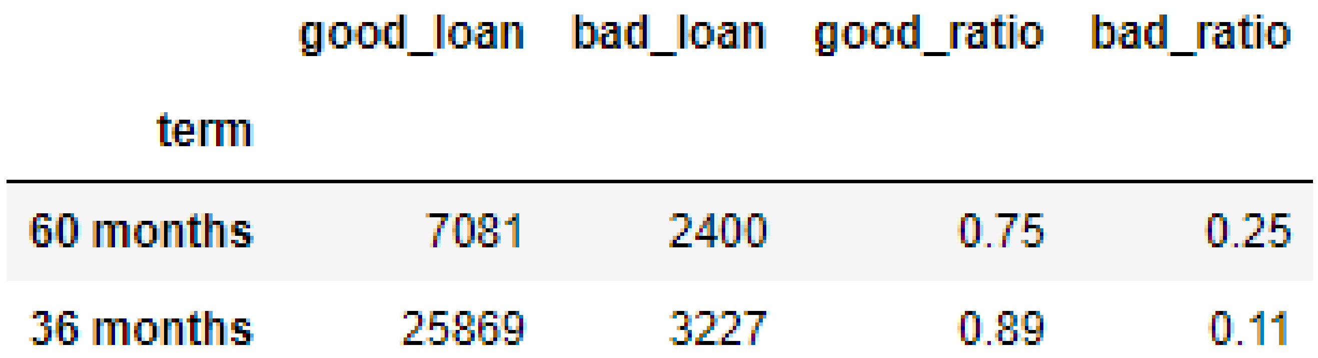
Summarize the outcome of each analysis

The analysis results provide the company with guidelines for each variable and its relation to the default probability. All visualizations presented the Bad Loans to provide clear vision of Default Probability.

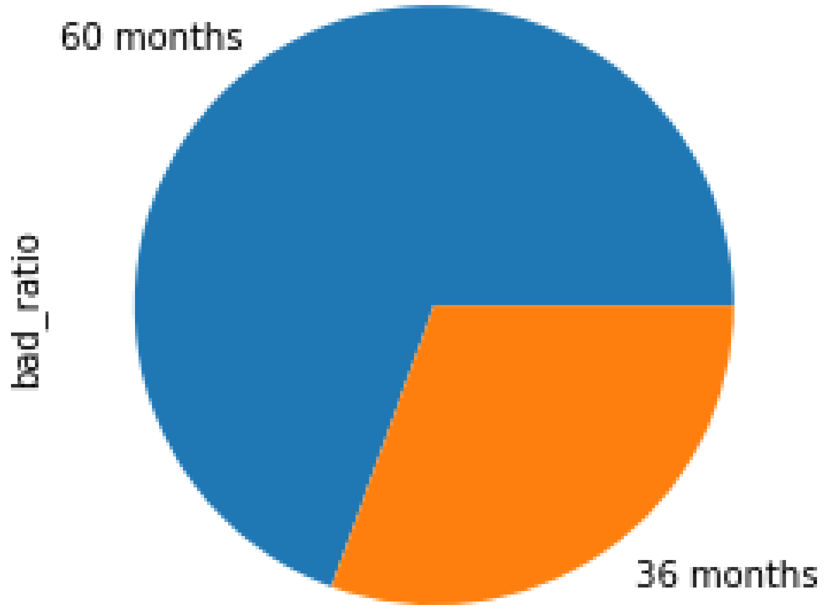
Sample Results – Term Analysis

Term Analysis: Term Analysis shows that 60 months terms loans has much higher probability of default. (25% vs 11% for 36months)

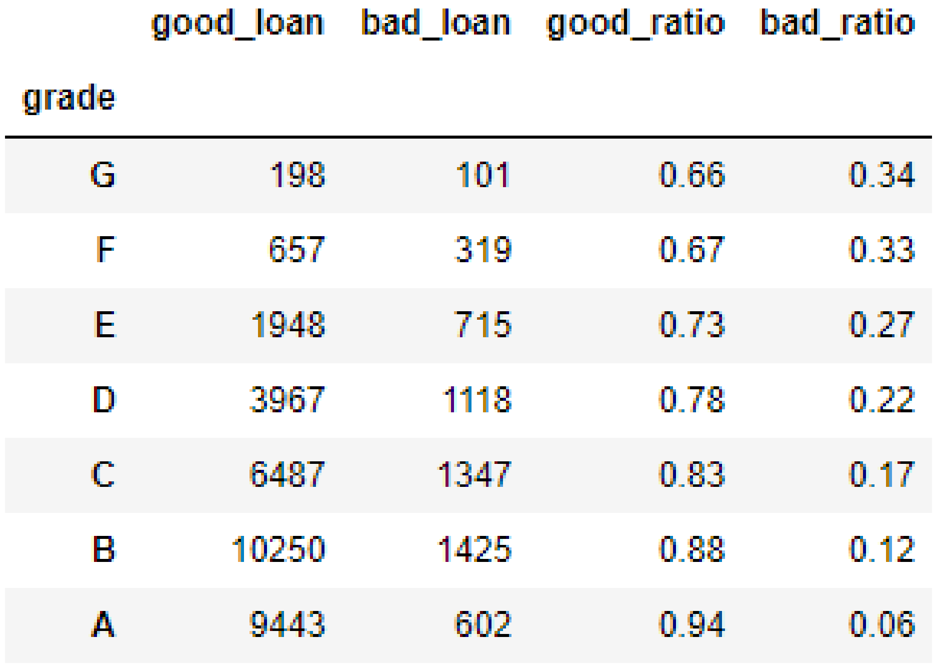
The company strategy should be to promote and focus on short term loans.



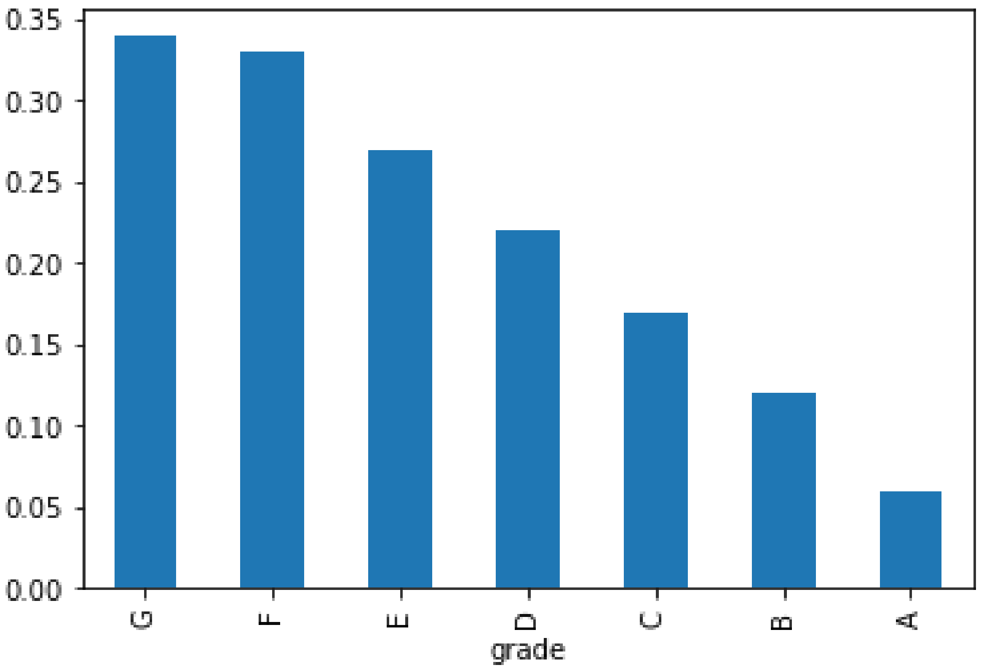
	good_loan	bad_loan	good_ratio	bad_ratio
term				
60 months	7081	2400	0.75	0.25
36 months	25869	3227	0.89	0.11



Sample Results – Grade Analysis

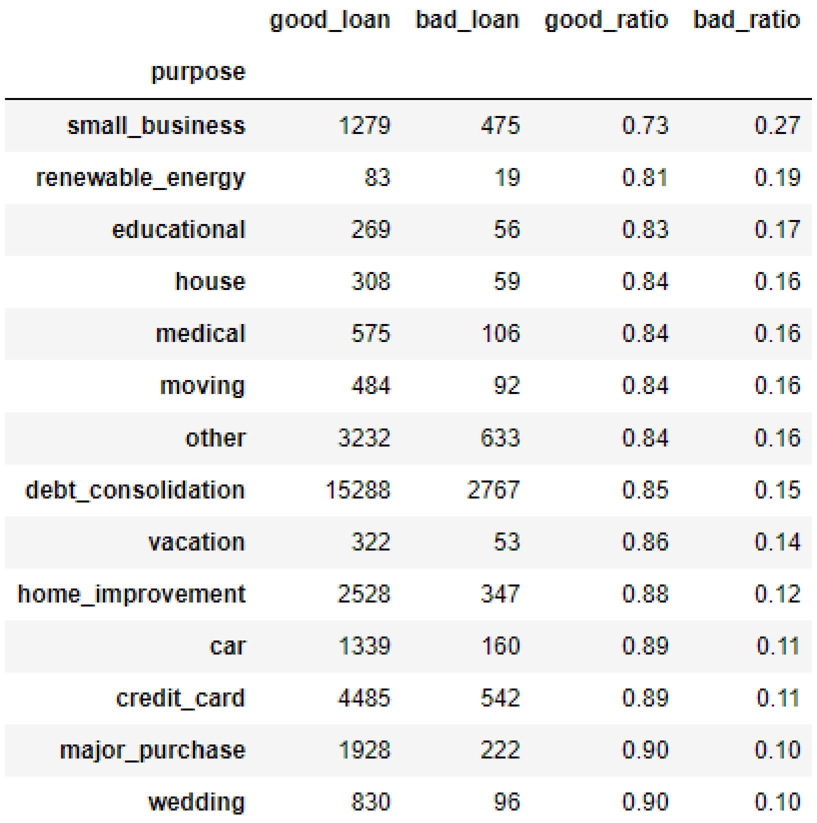


	good_loan	bad_loan	good_ratio	bad_ratio
grade				
G	198	101	0.66	0.34
F	657	319	0.67	0.33
E	1948	715	0.73	0.27
D	3967	1118	0.78	0.22
C	6487	1347	0.83	0.17
B	10250	1425	0.88	0.12
A	9443	602	0.94	0.06

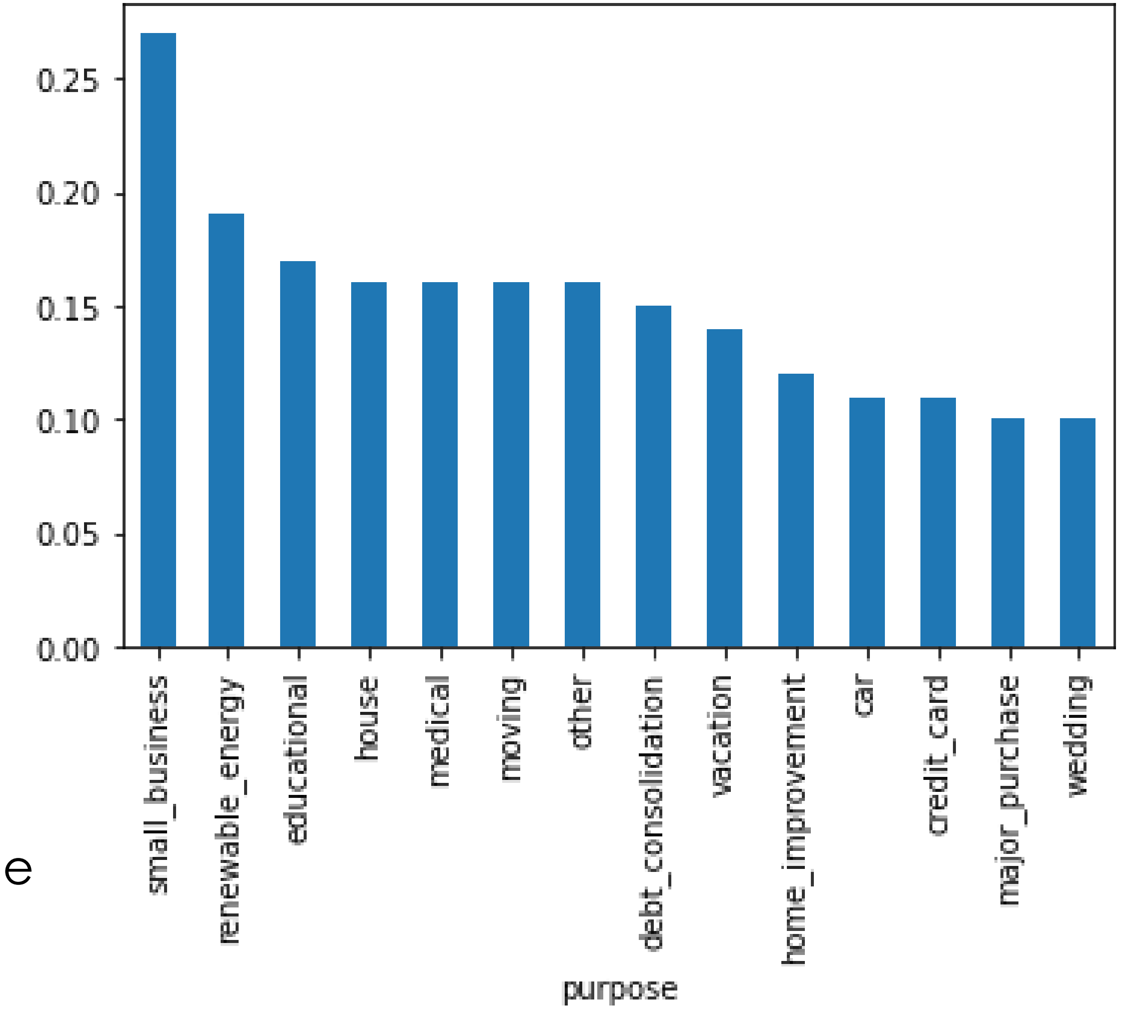


Grade Analysis shows that Default probability increases as we go from Graded A-B…G.
The company should focus on loans of grades with less bad ratio.

Sample Results – Purpose

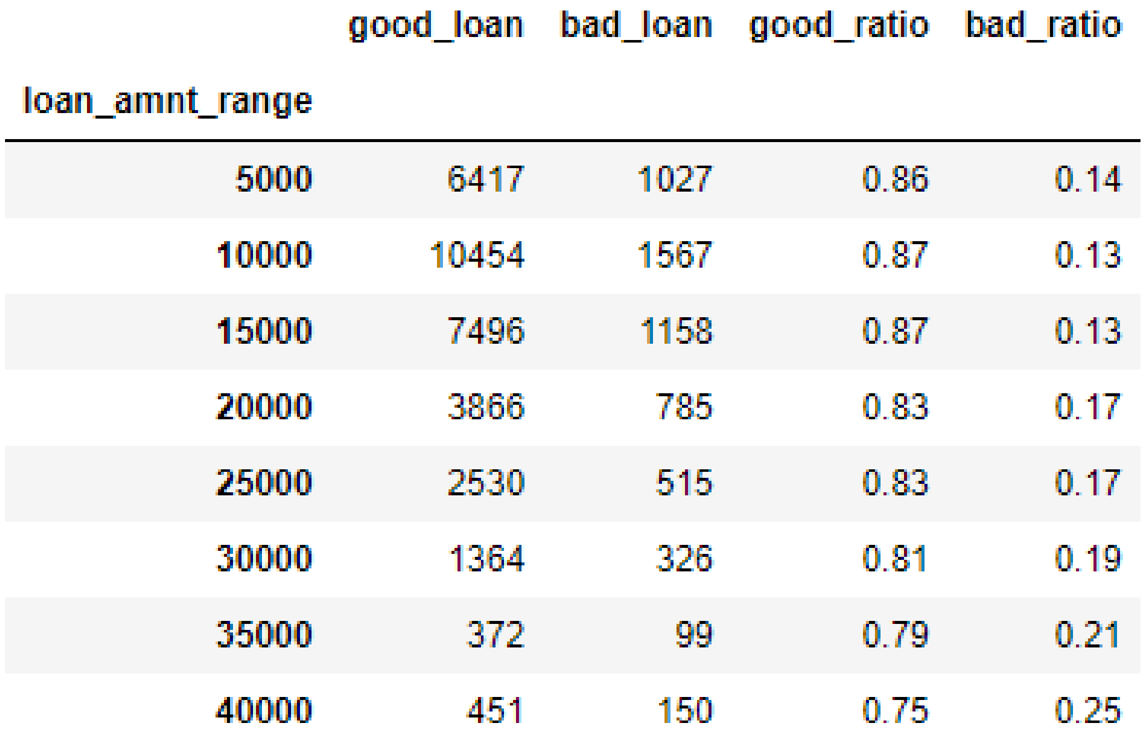


purpose	good_loan	bad_loan	good_ratio	bad_ratio
small_business	1279	475	0.73	0.27
renewable_energy	83	19	0.81	0.19
educational	269	56	0.83	0.17
house	308	59	0.84	0.16
medical	575	106	0.84	0.16
moving	484	92	0.84	0.16
other	3232	633	0.84	0.16
debt_consolidation	15288	2767	0.85	0.15
vacation	322	53	0.86	0.14
home_improvement	2528	347	0.88	0.12
car	1339	160	0.89	0.11
credit_card	4485	542	0.89	0.11
major_purchase	1928	222	0.90	0.10
wedding	830	96	0.90	0.10

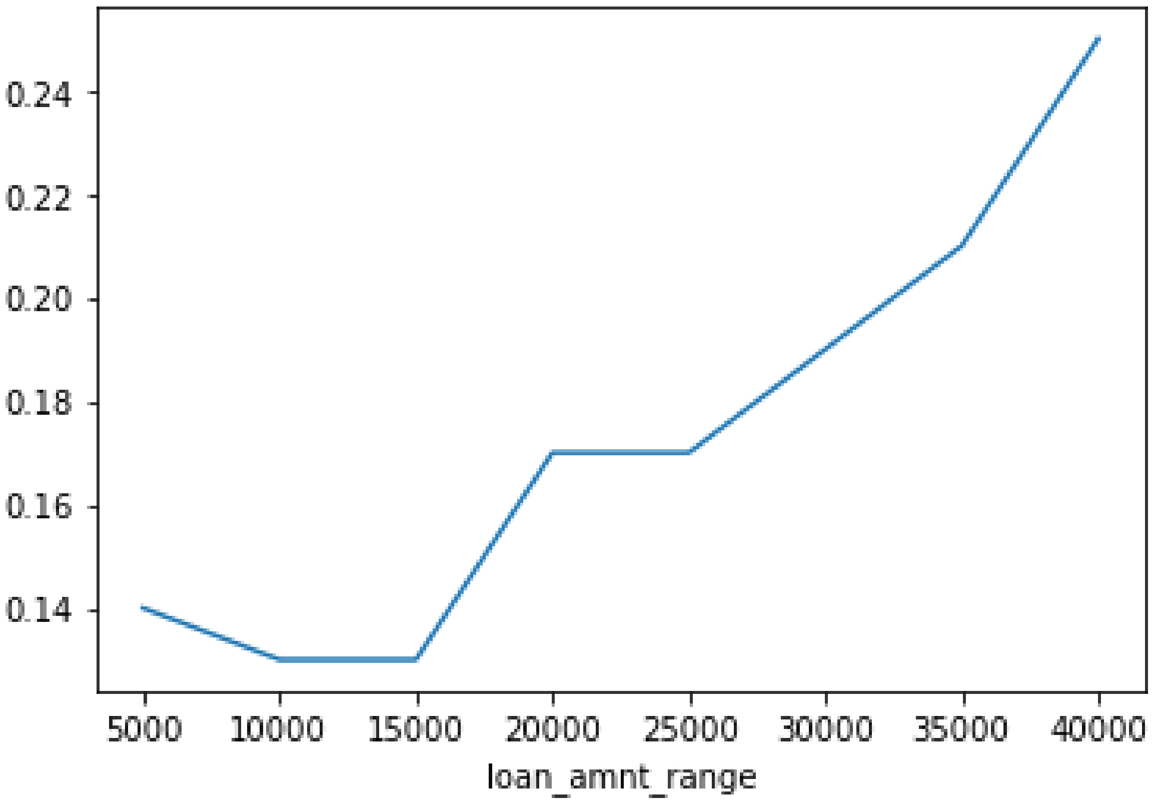


Small Business Loans have highest risk of default, while wedding are least risky. Avoid Small Business Loans

Sample Results – Loan Amount

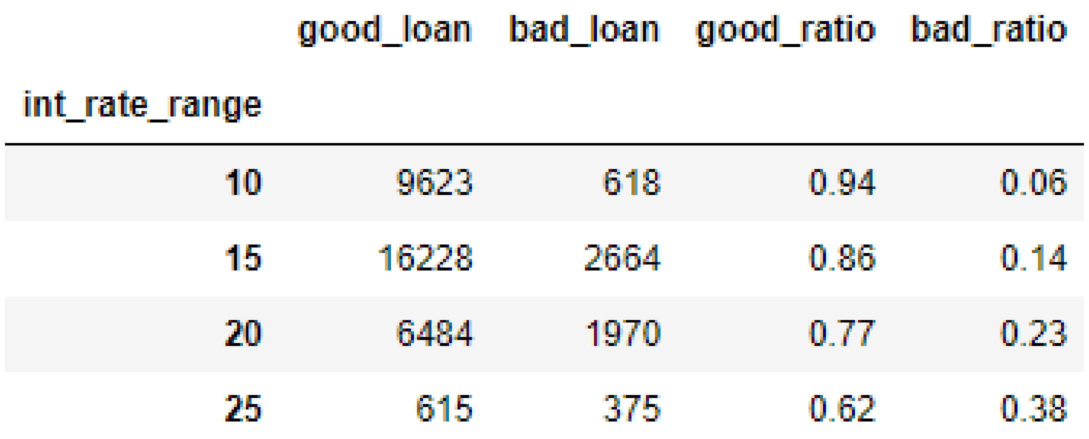


	good_loan	bad_loan	good_ratio	bad_ratio
loan_amnt_range				
5000	6417	1027	0.86	0.14
10000	10454	1567	0.87	0.13
15000	7496	1158	0.87	0.13
20000	3866	785	0.83	0.17
25000	2530	515	0.83	0.17
30000	1364	326	0.81	0.19
35000	372	99	0.79	0.21
40000	451	150	0.75	0.25

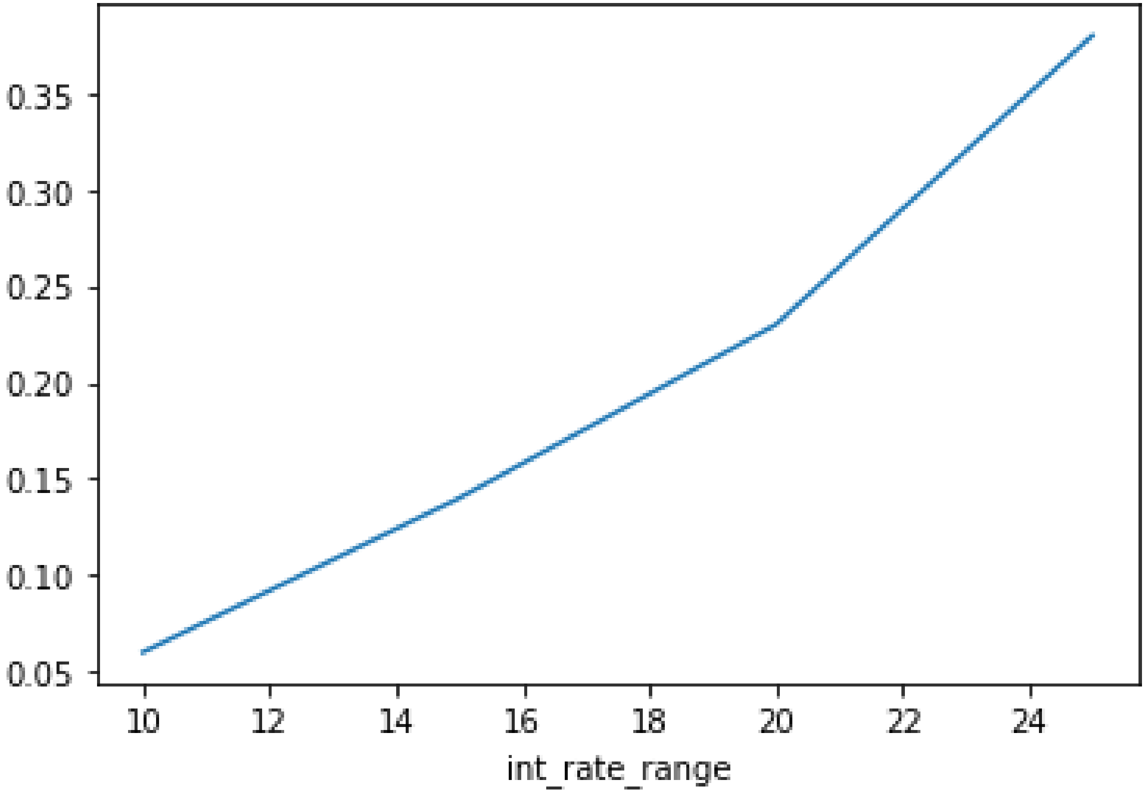


Loan Amount shows that default risk increases with higher amounts. The company must avoid high value loans.

Sample Results – Interest Rate



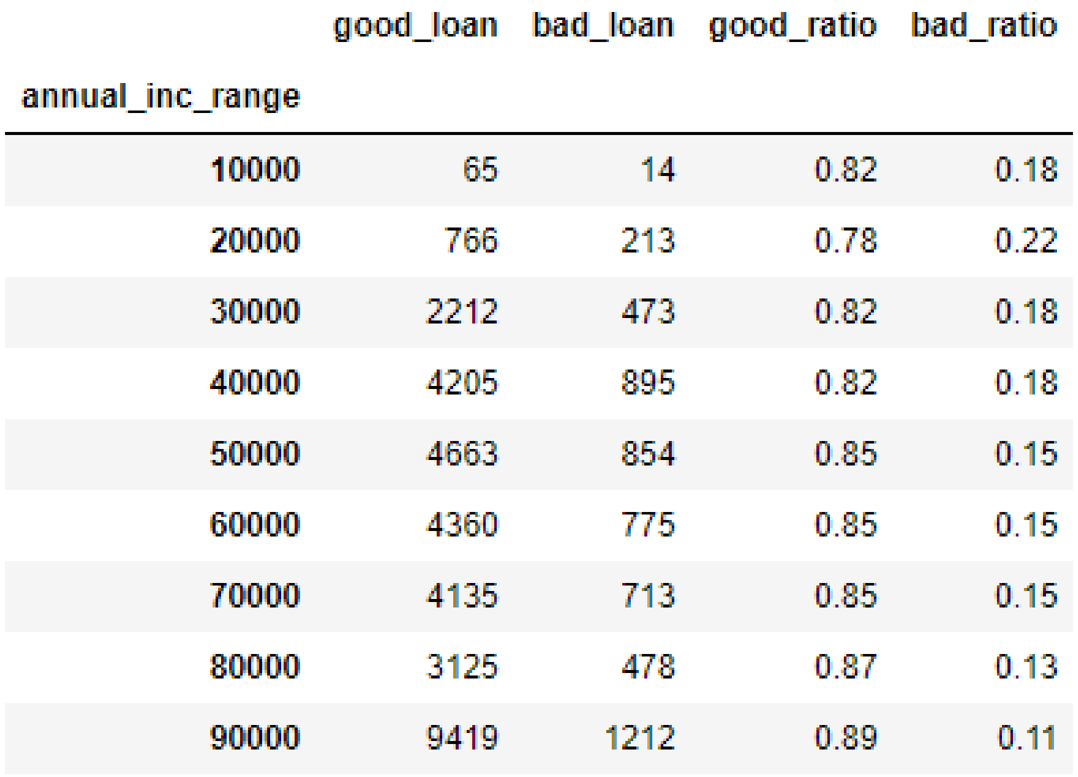
int_rate_range	good_loan	bad_loan	good_ratio	bad_ratio
10	9623	618	0.94	0.06
15	16228	2664	0.86	0.14
20	6484	1970	0.77	0.23
25	615	375	0.62	0.38



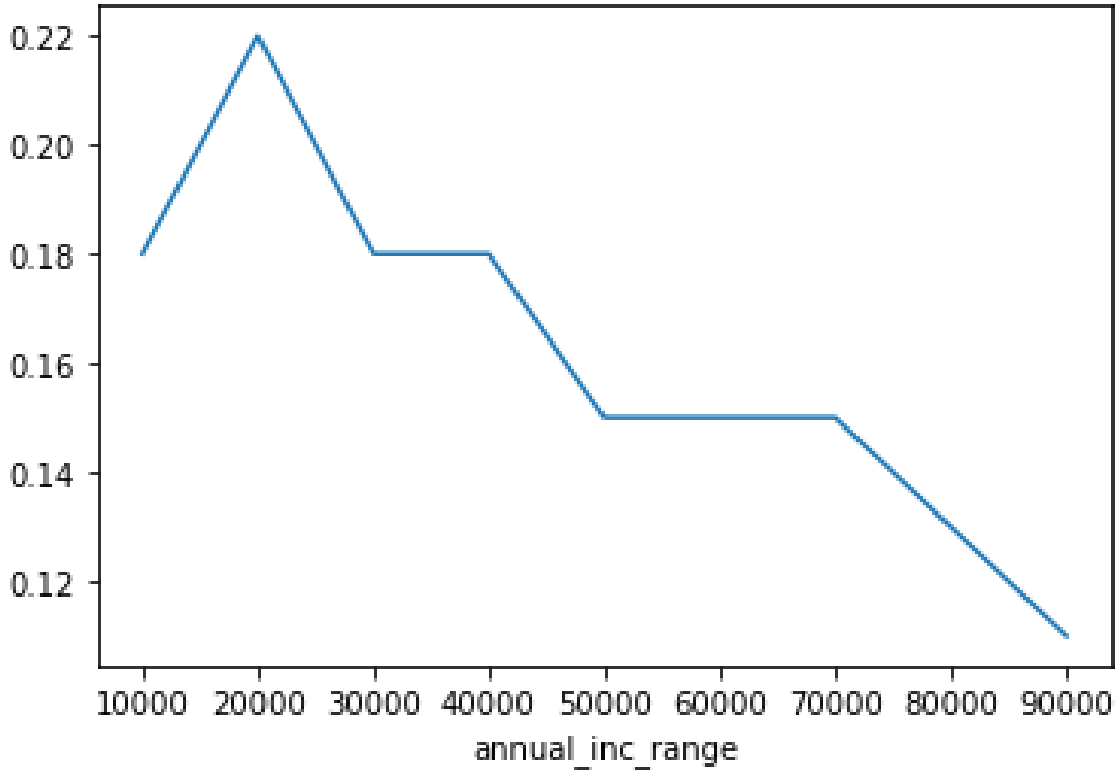
Loans with Higher Interest rate have higher default probability. The company must work out an equation between interest rate and default probability.

Higher Default probability leads to higher interest rate, and higher interest rate, leads to higher default probability.. Need a good Balance between the 2 indicators.

Sample Results – Purpose



annual_inc_range	good_loan	bad_loan	good_ratio	bad_ratio
10000	65	14	0.82	0.18
20000	766	213	0.78	0.22
30000	2212	473	0.82	0.18
40000	4205	895	0.82	0.18
50000	4663	854	0.85	0.15
60000	4360	775	0.85	0.15
70000	4135	713	0.85	0.15
80000	3125	478	0.87	0.13
90000	9419	1212	0.89	0.11



Annual income is another predictor; Low annual income is associated with higher Default probability. The company should avoid loans for applicants with very low annual income

Conclusion

We have provided sample results and recommendations from the analysis. For the complete Analysis on all variables, you can run the Python code.

For the Continuous Variables analysis, different bins can be tested when calling the function. We used bins that are logical with the variable, for example 5% for interest rate, 1000 for loans amount, and 10000 for annual income. Other bins can be tested for more analysis.

Some variables were found of no, or very little influence on the default probability like the employment length.

For the full analysis, please run the Python code.