LENDING CLUB CASE STUDY

SUBMISSION

Name: Srinivas soma

Gajula Jagadeesh

Lending Club Case Study (EDA)

This is case study on lending club which is a online portal where people apply for loan and borrow money. Lending club is for two kind of people. Lenders and borrows. We are on lenders side. We are provided with data and we need to conduct **Exploratory data analysis** on data. Data contains more than 100 variables but, We are concerned about **loan_status** the target variable which contains 3 categories of loan status. Fully paid, current and charged off. We need to tell the most effecting variables in an application and hints of loan application will be defaulted. So that lenders or investors will be aware of the coming events.

Loan approvals will be based on analysis conducted. So every hint and variable matters. This EDA eases application process even before it is shown to lender. Applications will be filtered in minutes and loan status will be predicted in real time.

Steps involved

UpGrad

Loan.csv

Importing
- Data importing
- Data understanding and observing

Basic cleaning
- Dropping all null columns
- Dropping cols with one unique values

More Cleaning
- Dropping cols with missing values > 50
- Examining rows

Advanced cleaning
- Imputing and cleaning rows
- Manually dropping less significant cols

Separation variables
- Creating dataframe having only categories
- Creating dataframe having numeric values

Concise
- Merging numerical columns to concise data
- Cleaning categorical variables like datetime

Univariate analysis
- Plotting numerical distributions
- Plotting categorical data

Univariate
- Observing numerical data
- Understanding categorical data

Bivariate analysis
- Grouping and plotting categorical variables
- Calculating significance for numerical data

Creating heat maps
- Heatmap for numerical values
- Heatmap with loan_status vs variables

Filtering numerical
- Filtering and taking only significant variables

Pair plot
- Pair plotting numeric correlations
- Plotting distributions on types of loans

Observations from data

Data understanding:

- Dataset contain 111 columns and approx. 39k rows
- More than 50 columns were complete null values
- About 14% of loans were defaulted
- Few numerical columns distributions where skewed

Basic cleaning:

- 57 were left after dropping all null columns
- 6 columns were having only one unique value so dropped since it can’t help us
- 3 columns were found with more than 50% of missing values
- maximum values missing in any given row is 6

Advanced Cleaning:

- Need to drop irrelevant columns like id, member_id, since they are purely random
- Since dataset is large enough and we can confidently drop few rows with missing values
- Separating numerical and categorical data is easy way

Observations from data

Separation of numerical and categorical data:

- 23 columns have pure numerical data
- 13 were categorical (datatype object may contain datetime and percentage values)
- loan_amnt, funded_amnt and funded_amnt_inv distributions were identical
- 4 columns contained datetime

Univariate analysis:

- More numerical column distributions were skewed
- 8 columns were very crucial

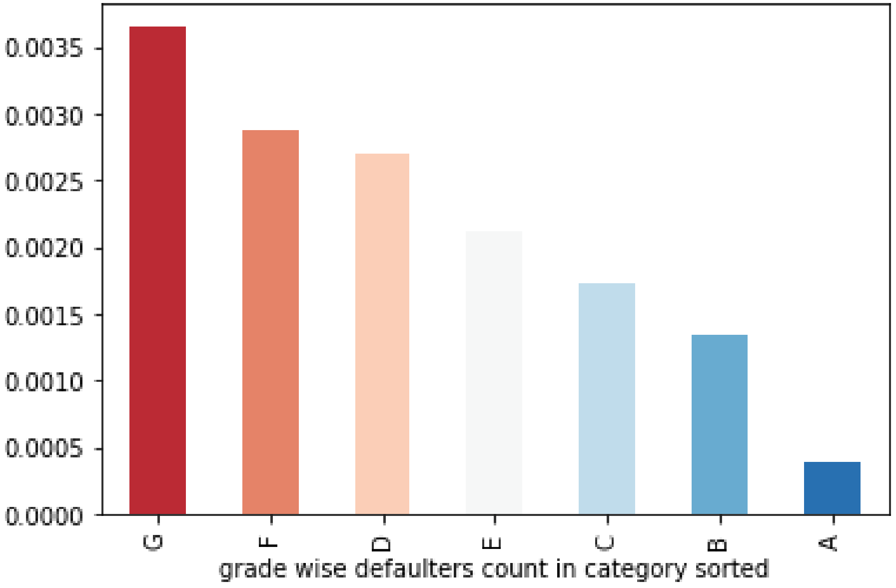
Bivariate analysis:

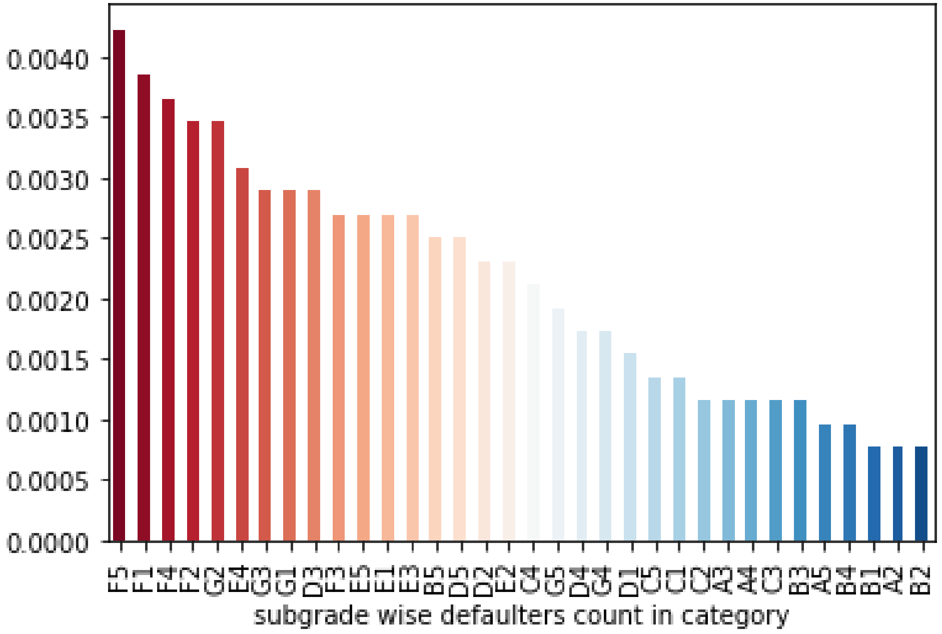
- Data is grouped based
- Risk factors are discussed further in PPT

Numerical Analysis:

- lot of numerical variables were found insignificant

Categorical analysis and hints





Variable grade:

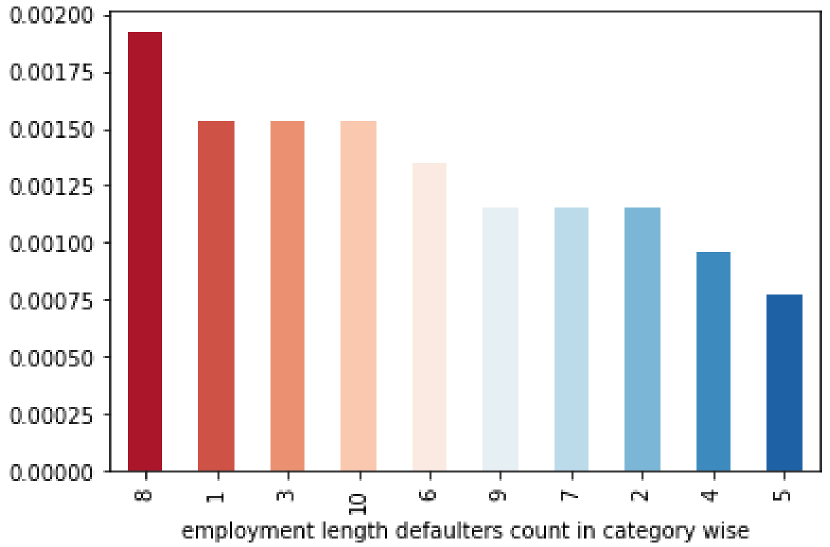
G grade have highest defaulters followed by F grade

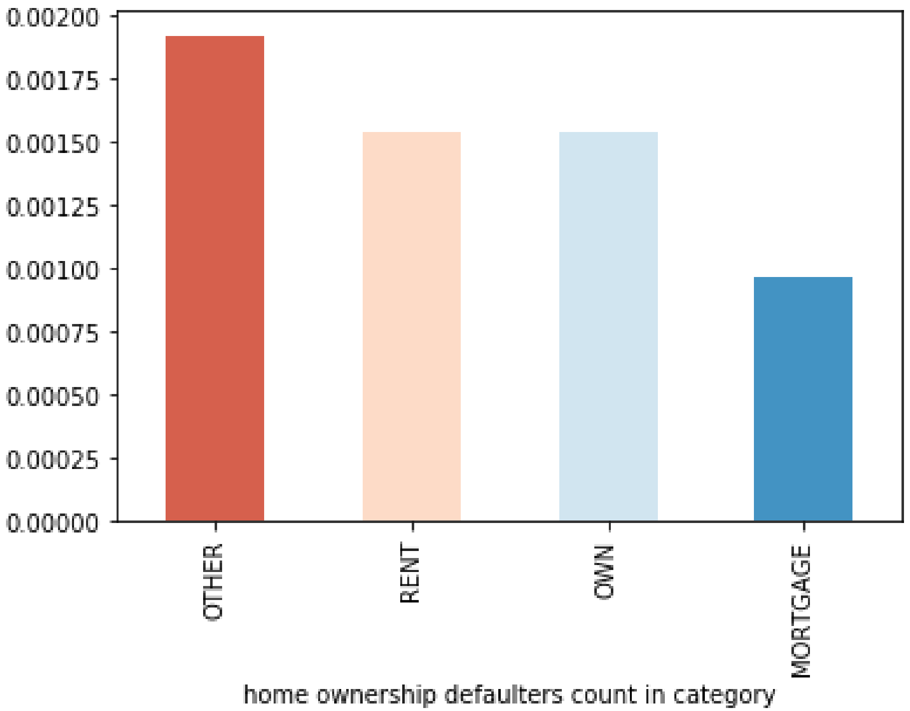
Variable sub_grade:

F5 have highest defaulter rate

sequence risk is given above

Categorical analysis and hints





Category emp_length:

emp_length > 8 and < 4 is risky

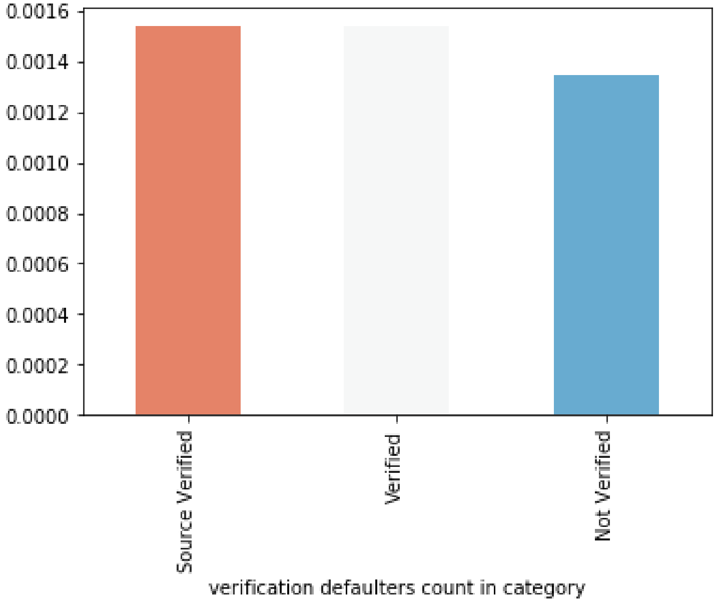
8 and 1 year have more defaulters

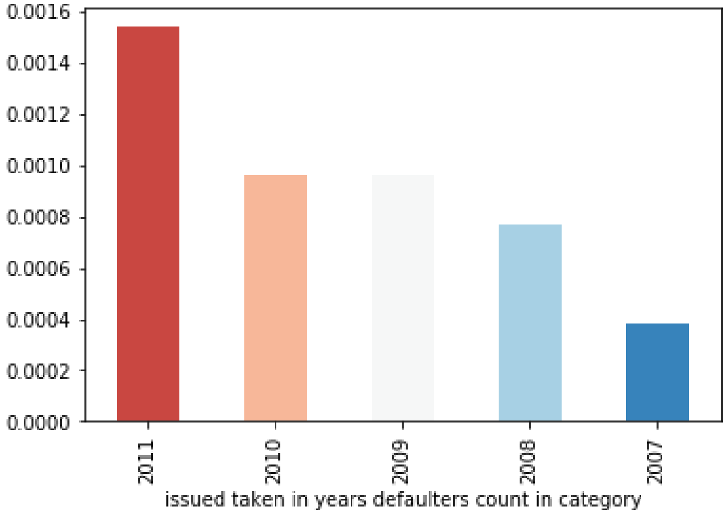
Category home_ownership:

Other category is more risky

rent and own have same level risk so not a problem

Categorical analysis and hints





Category verification status:

Verification status don't have much significance

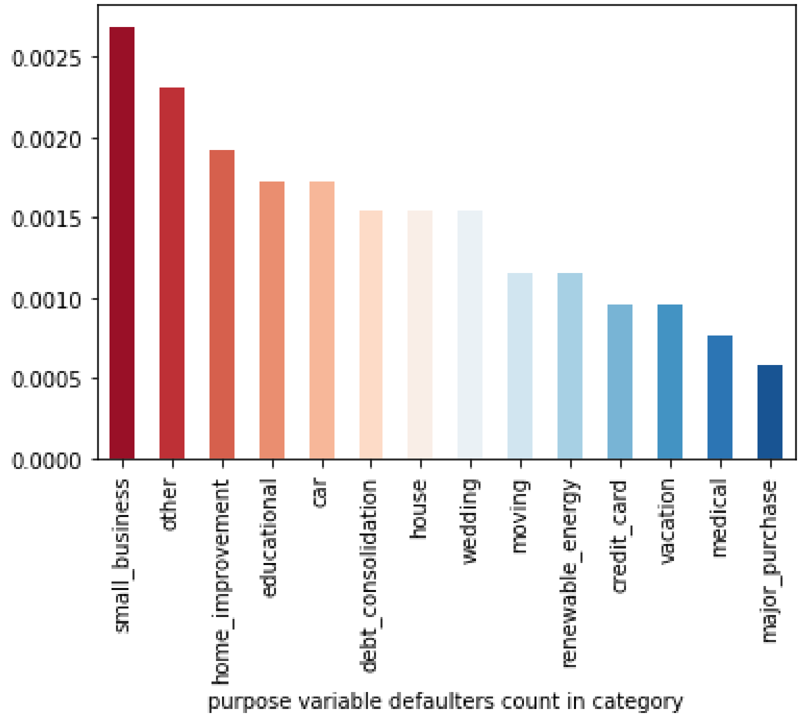
Category issue_d_year:

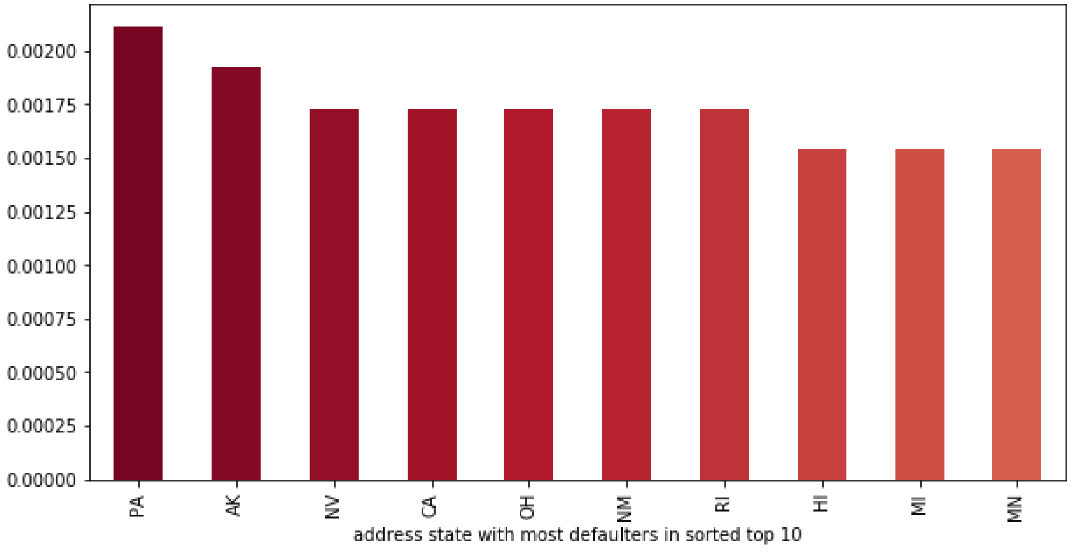
-this is derived from issue_d column

-2011 have highest defaulter count.

-defaulters were steadily increasing from 2007 to 2011

Categorical analysis and hints





Category purpose:

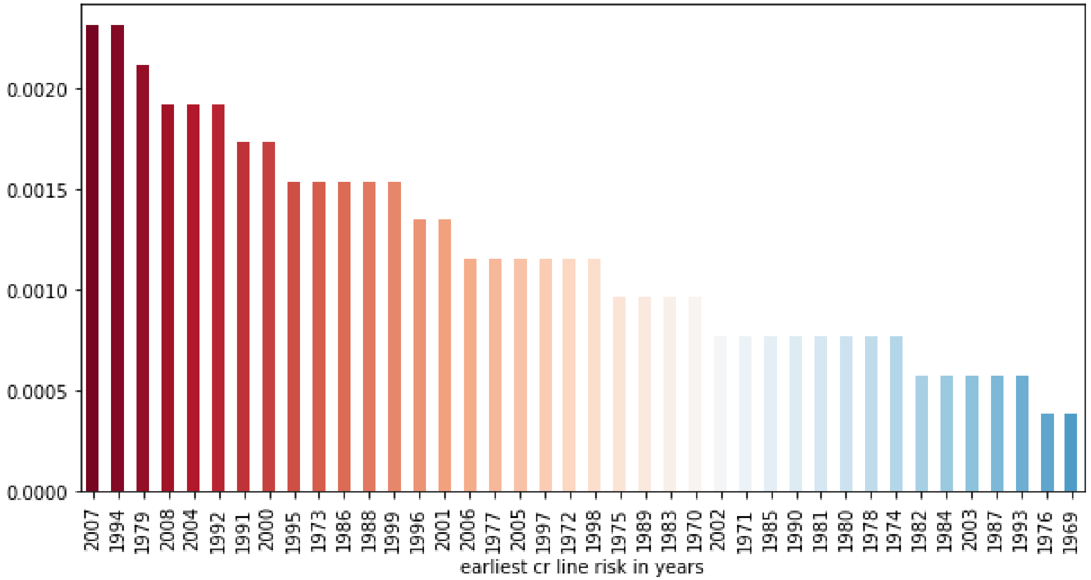
Top 10 states with high defaulter rate is plotted above

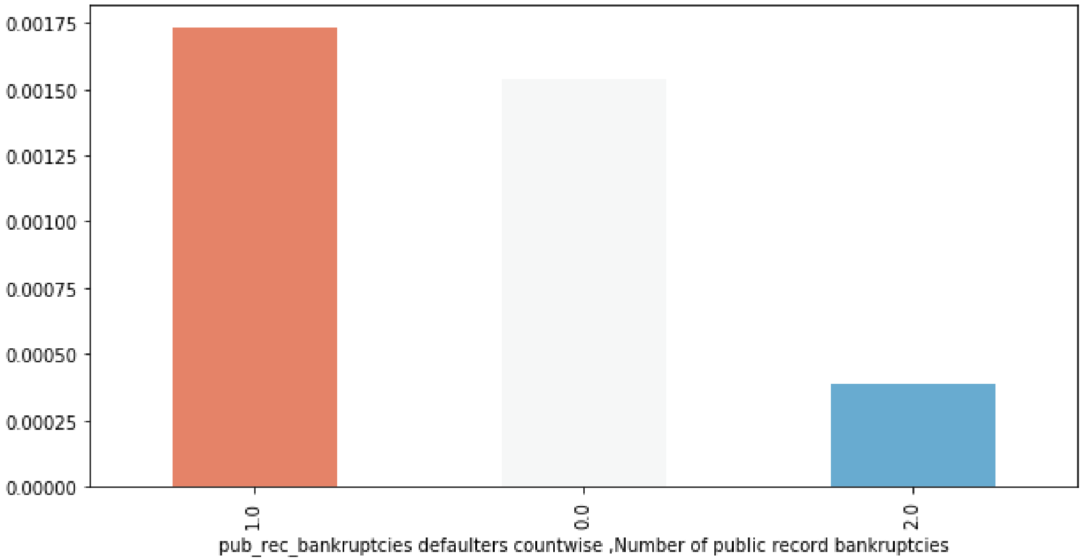
Top states were PA and AK

Category purpose:

small business have highest defaulting risk followed by other

Categorical analysis and hints





Category earliest_cr_line_year:

this is a derived metric

2007 and 1994 have highest defaulter count

Category Pub_rec_bankruptices:

here more count is on one record of bankruptcy

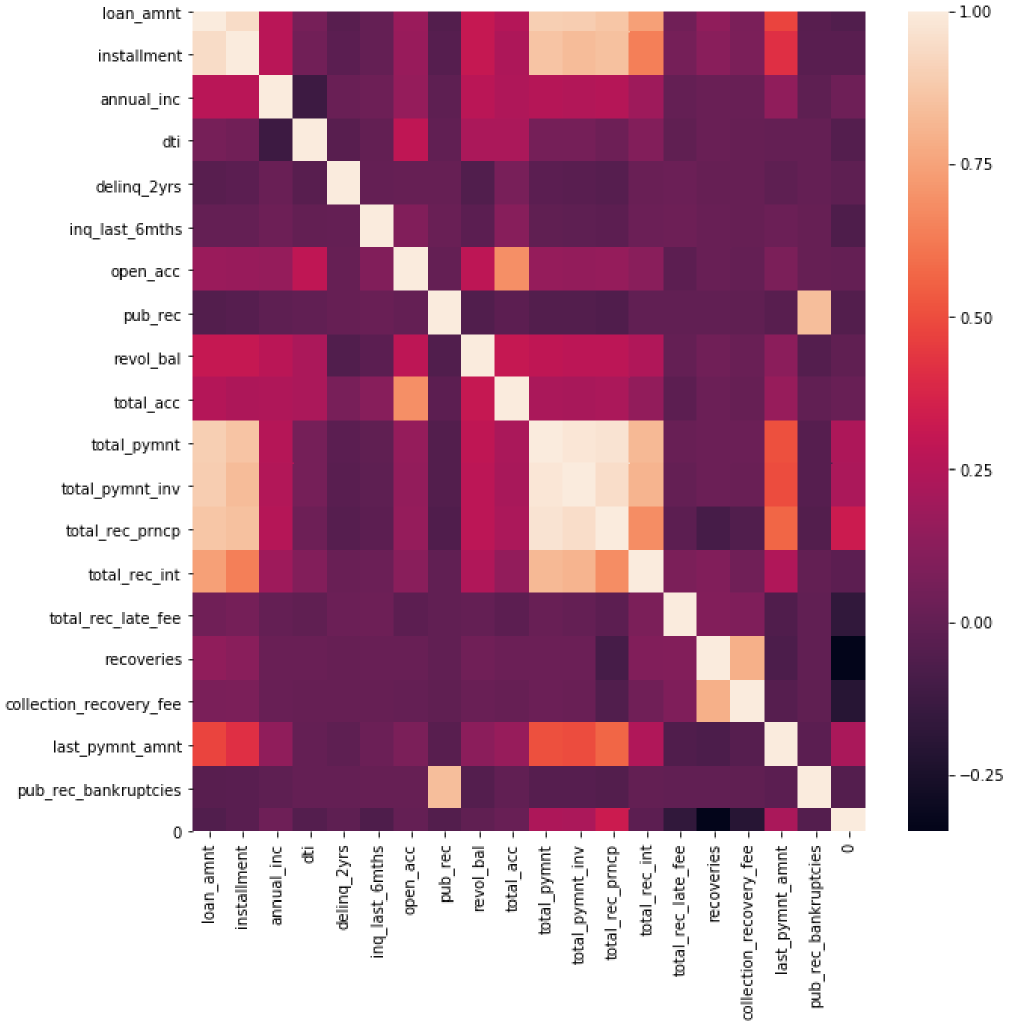
its doesn't state any risk in sequence, But not recommended to approve loan easily with at least one bankruptcy

Numerical variables analysis

- Not all the given variables have significant effect on loan_status. This is calculated by feature importance method with sklearn ensemble classifier.
- Significant numeric variables have effect on outcome are

loan_amnt
installment
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt

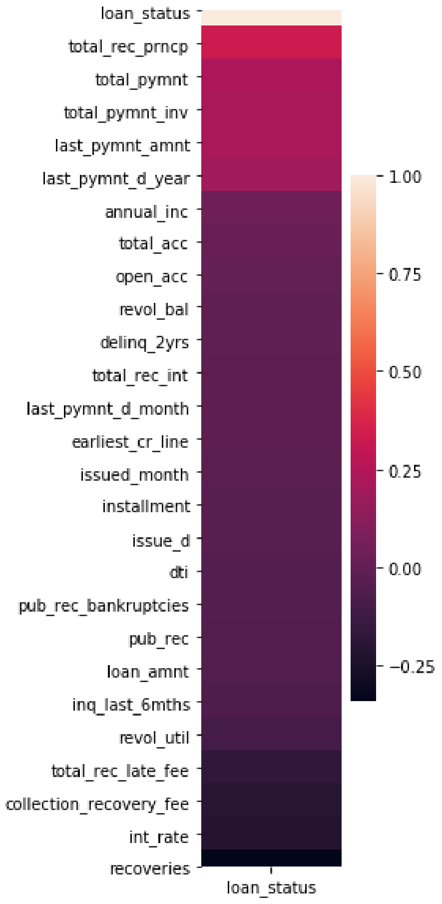
Numerical analysis



This is correlation heatmap of all numerical variables

- Loan_amnt and installement have good correlation
- Total_payment , total_payment_inv and total_rec_princp have good correlation
- Last ZERO indicates loan_status have a strong negative correlation with recoveries
- Lot other insights can be drawn from this single picture

Numerical analysis



This is correlation heatmap loan_amnt with all other variables.

- Lot of variables have approx. zero and negative correlation
- Only few variable share positive correlation.
- You can see heatmap and conclude the which variables could lay intuition for outcome

plot

Pair plot between all significant

Numeric variables

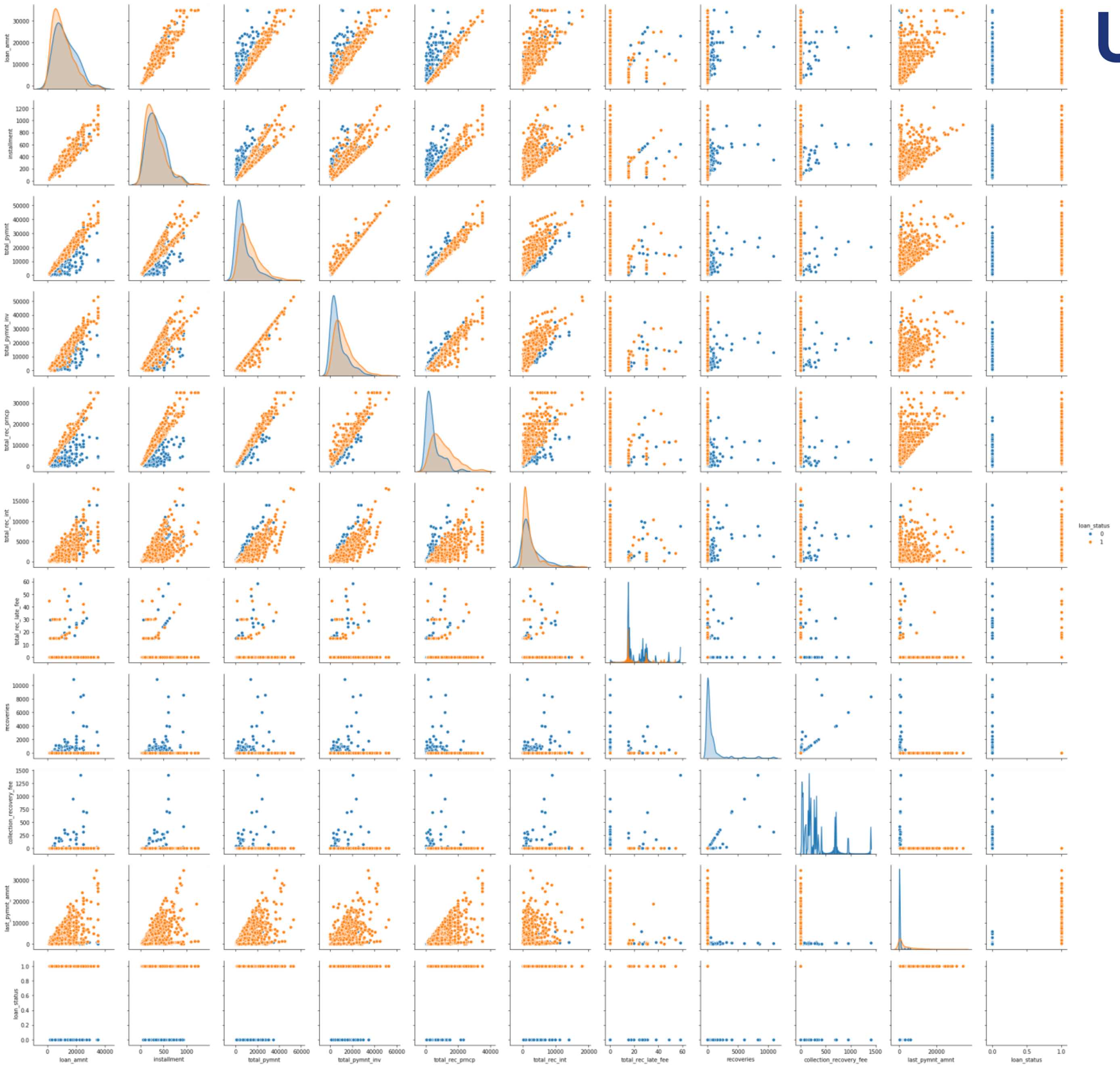
- You can see correlation in plots
- We actually trained a classifier

model only on numerical data

built with keras and got upto 95+

accuracy.

[Github link]



Conclusion

Lending club data is real world data. Every insight help lender for better outcome and its really intuitive for learners. Numerical was bit confusing because distribution is not even. Categorical data was little easier.

This Case study is really a great experience and exposure to real world datasets. we got to learn about risk analysis for the first time. Exploratory data analysis have no end we can each into each variable and draw new insights. We are just limited with time and resources. We don’t need that much of depth in every aspect, think it will go in logistic fashion and flat out at some point. Since too much preparation won’t work. we wanted to do further analysis with derived metrics. We got to learn some profound things and sanity ways of exploring into datasets. I really thankful for group mate. He was guiding me with better suggestions and same in case of student mentor. Its really a great hands on experience.

Thank you