EDA Loan Case Study

BY ABHINAV AGGARWAL

Content

Introduction

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business Objectives

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Data Understanding

- This dataset has 3 files as explained below:
- loan.csv: which contain complete loan data for all loans issued through the time period 2007 t0 2011.
- Data_Dictionary.xlsx:which describes the meaning of these variables

Data Cleaning part1

- Check which all column has data as empty or NAN are more than 30%(so that we can remove those column)

  Answer:

```
Index(['desc', 'mths_since_last_delinq', 'mths_since_last_record',
       'next_pymnt_d', 'mths_since_last_major_derog', 'annual_inc_joint',
       'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal',
       'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m',
       'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m',
       'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi',
       'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal',
       'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct',
       'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl',
       'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq',
       'mths_since_recent_inq', 'mths_since_recent_revol_delinq',
       'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl',
       'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl',
       'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m',
       'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m',
       'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim',
       'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit'],
      dtype='object')
```

- Remove these column
- Remove unwanted column like identification column and column with incomplete data example zip code

Data Cleaning part11

- Fill empty spaces with with not available in text column

```
print(Column)
#repcaling empty places in emp_title','emp_length','title','last_pymnt_d','last_credit_pull_d' with Not avaliable
Loan_DS[['emp_title','emp_length','title','last_pymnt_d','last_credit_pull_d']] = Loan_DS[['emp_title','emp_length',
Column = Loan_DS.isnull().mean()*100
```

- Find the correct value to fill in numeric column like mode or median

```
Mod_pub_rec_bankruptcies=Loan_DS["pub_rec_bankruptcies"].mode()[0]
print(Mod_pub_rec_bankruptcies)
Mod_tax_liens=Loan_DS["tax_liens"].mode()[0]
print(Mod_tax_liens)
Mod_collections_12_mths_ex_med=Loan_DS["collections_12_mths_ex_med"].mode()[0]
print(Mod_collections_12_mths_ex_med)
Mod_chargeoff_within_12_mths=Loan_DS["chargeoff_within_12_mths"].mode()[0]
print(Mod_chargeoff_within_12_mths)


Loan_DS['collections_12_mths_ex_med'] = Loan_DS['collections_12_mths_ex_med'].fillna(Mod_collections_12_mths_ex_med)
Loan_DS['pub_rec_bankruptcies'] = Loan_DS['pub_rec_bankruptcies'].fillna(Mod_pub_rec_bankruptcies)
Loan_DS['tax_liens'] = Loan_DS['tax_liens'].fillna(Mod_tax_liens)
Loan_DS['chargeoff_within_12_mths'] = Loan_DS['chargeoff_within_12_mths'].fillna(Mod_chargeoff_within_12_mths)
Column = Loan_DS.isnull().mean()*100
print(Column)
```

Data Cleaning part111

- Removing those column which has same values

```
print(Loan_DS.shape)
Loan_DS.nunique()
nunique = Loan_DS.nunique()
cols_to_drop_nunique = nunique[nunique == 1].index
print(cols_to_drop_nunique)
#
Loan_DS=Loan_DS.drop(cols_to_drop_nunique, axis=1)
print(Loan_DS.shape)
```

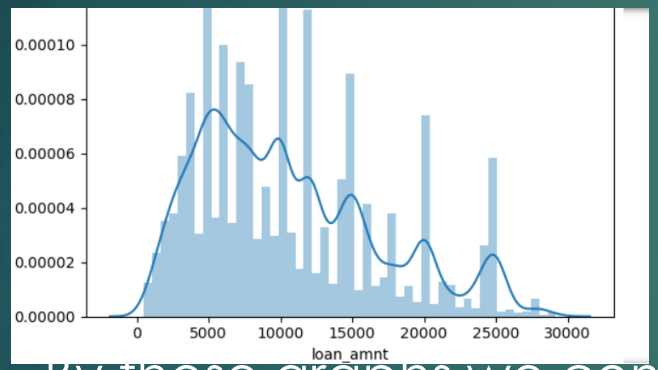
```
(39717, 49)
Index(['pymnt_plan', 'initial_list_status', 'collections_12_mths_ex_med',
       'policy_code', 'application_type', 'acc_now_delinq',
       'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens'],
      dtype='object')
(39717, 40)
```

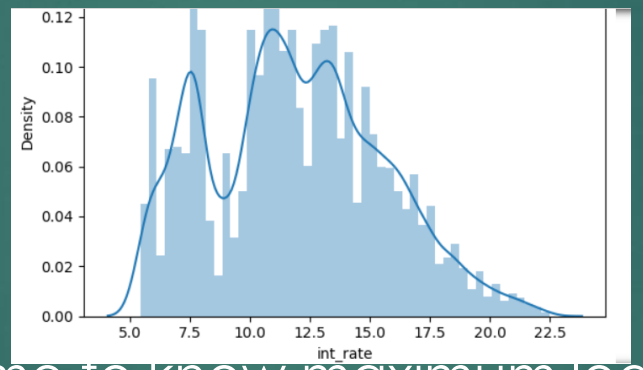
Data Cleaning part IV

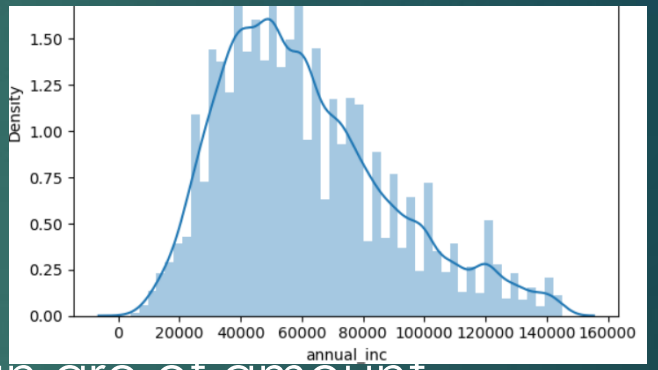
- Converting column to numeric type and removing % from interest rate column
- Removing outliner in loan amount ,interest rate and annual salary
- Removing the outliner based on purpose of loan those loan purpose are less than 0.5% and others (renewable_energy as loan purpose for this was around 0.25%) and remove based on house ownership like none and others which are not useful for analysis

Data Analysis I

- Create list of continuous and categorical columns
- Plot graph for Continuous columns to understand behaviors







- By these graphs we come to know maximum loan are of amount 5000 to 15000 with interest range of 8.9% to 14.3% and person who has taken these loans has annual income in the range of 40000 to 77000 USD
- Please check Box plot too in the code for exact values

Data Analysis II

- Total loans account under analysis are 32923 and out of which 83%(approx.) are already paid off and 13%(approx.) are charged off and remaining are currently running
- Maximum loans are taken from CA region
- Maximum Loans are for 36 months
- Maximum loans taker are living on rent
- Maximum persons have experience more than 10 year and loans are not verified
- Maximum loans which are taken on the purpose of moving are less chances of charged off
- Currently only 60 months loans are running