EDA – Lending Club Case Study

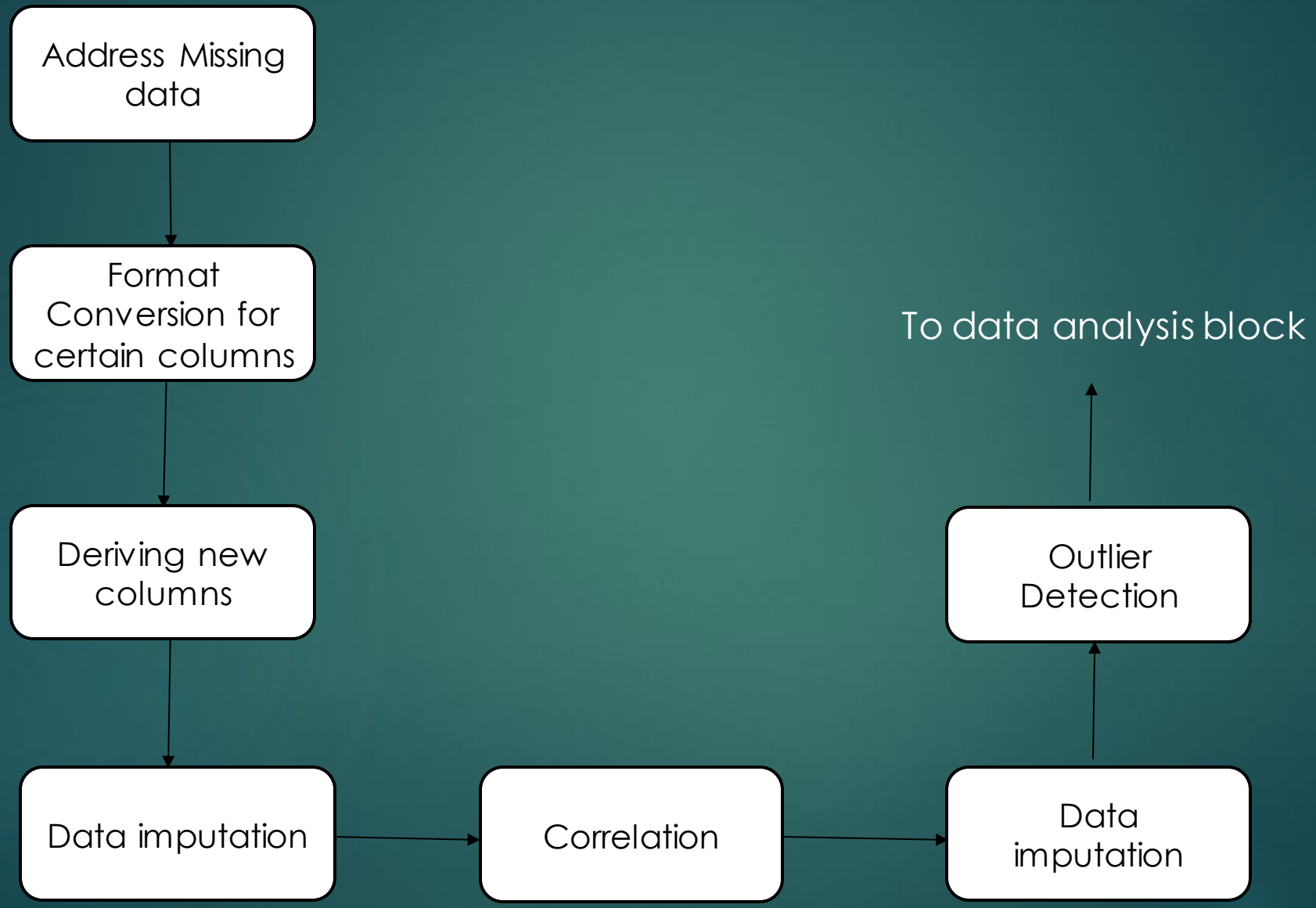
SANTOSH KRISHNAN

EPGP MAY COHORT IN AI & ML

Approach taken - Overview

- The entire process that I've taken can be split into two blocks viz. *Data Cleaning* and *Data Analysis*.
- *Data Cleaning* addresses all issues with the columns and rows. It manipulates them and brings them in to suitable formats, if needed. New variables are also derived to extrapolate more information from the data.
- *Data Analysis* does both univariate and bivariate analysis to identify patterns, correlation, etc. This block is where most of the driver variables are identified.
- This presentation is structured in the aforementioned sequence. Some observations may be skipped here to avoid redundancy, but all of the observations are present in a detailed manner in the jupyter notebook.

Data Cleaning blocks involved



Addressing Missing Data Block

- Columns with more than 60% missing data are removed.
- Special Cases: id, member_id, url are not features. Therefore, these columns are removed. They do not give any information, hence they are dropped from analysis as well.
- *Total_xx* columns: Some of these columns have information that gets filled up after loan is started. But we will keep them anyway to hopefully identify some patterns.

Format Conversion for certain columns block

- Term: Converting this to numeric and then convert to years instead of months.
- Int_rate: Removing % symbol
- Emp_length: Removing the string "years" and then convert it into buckets: 0-2, 2-4, 4-6, 6-8, 8-10 and > 10 years.
- Revol_utiL: Removing % symbol

Quasi Constant Variables block

- Columns with more than 95% constant data are removed.

Deriving new columns block

- Annual_inc: Categorizing this column in to bins of $20,000.
- Funded_amount: Categorizing this column in to bins of $5,000.
- Int_rate: Categorizing this column in to [0-8, 8-10, 10-12, 12-14, 14-16, >16]
- From columns, *issue_d* and *last_payment_d,* we compute the time elapsed from the issue date in years(float) i.e. if 5 years and 3 months have elapsed, it will be represented as 5.25 years. Column name: delta.
- This is then categorized in to [0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, 7-8]

Data imputation block

- Columns, *title, revol_util, last_payment_d, last_credit_pull_d, emp_length and pub_rec_bankruptcies* are filled with their mode values.

Correlation block

- A heatmap is plotted for the correlation matrix and only one variable of a strongly correlated group is kept and others are discarded.
- Eg: loan_amnt, funded_amnt, funded_amnt_inv are strongly correlated. So we drop loan_amnt, funded_amnt_inv and only keep funded_amnt.
- We could have easily done our analysis with loan_amnt and dropped funded_amnt instead and our observations wouldn't have changed, since

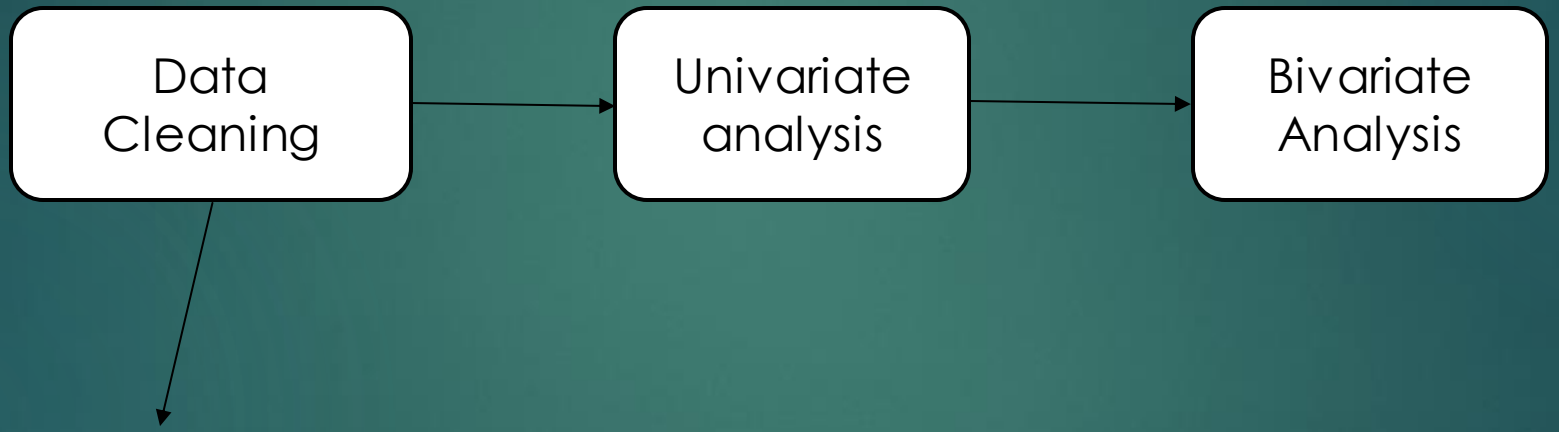
  funded_amnt <= loan_amnt.
- Following metric is used to determine strength of correlation:
  - $|r| < 0.3$ ⇒⇒ None or Very Weak
  - $0.3 < |r| < 0.5$ ⇒⇒ Weak
  - $0.5 < |r| < 0.7$ ⇒⇒ Moderate
  - $|r| > 0.7$ ⇒⇒ Strong

Outlier Detection block

- We use box plots and IQR analysis to detect outliers.
- For this data, however, many rows in the top 1% seem to have charged off loans, fully paid and currently running loans.
- This is observed across all columns.
- Dropping all of these rows will shrink the data drastically, therefore, we retain them because it might contain important information needed for analysis.

Data Analysis Blocks

Once data cleaning is done, we can move to data analysis section



Univariate Analysis block

- We first split the columns in to categorical and continuous.
- We then use box plots of every continuous column with hue on loan_status and obtain inferences.
- We then plot probability distribution of certain columns with seaborn's displot.
- The observations obtained from univariate analysis is mentioned in the section III.1.1. Box Plots of III.1. Univariate analysis in the notebook.

Bivariate Analysis block

- We already looked at correlation block, now we will do bivariate analysis with a keen focus on loan_status.
- Following columns are compared against loan_status:
  - Grade: Borrowers with bad credit ratings tend to borrow more money and they tend to default more often.
  - Home Ownership: Could not identify a good pattern.
  - Verification Status: Could not identify a good pattern.
  - Purpose: Borrowers borrowing money for Small business tend to default more often than other purposes.
  - Emp_length(experience): Could not identify a good pattern.
  - Annual Income: Lower income borrowers are defaulting more often than higher income borrowers.
  - Funded amount: Higher the amount, more is the charged off percentage

Bivariate Analysis Block(cont.)

- Rate of interest: Higher interest rates, attract more defaults.
- Public records: Could not identify a good pattern.
- Inq_last_6mths: Borrowers who have inquired more in the past 6 months are more likely to default
- State address: Borrowers from the state of Nebraska(NE) are more likely to default than others.
- Loan Term: 5 year loans have more defaults than 3 year loan terms. Also interest rate is higher for 5 year loans.
- Delta_bins(time elapsed from loan issue date): Percentage of payment received in conjunction with delta bins give us a good indicator of potential loan defaults. Since the bank only has two terms, it is best to check at the term boundaries. i.e. For the 3 year loan, check bin = 2-3 and percent_payment_bin = 0-80

Bivariate analysis in between other columns(minus loan_status)

- Grade vs interest Rate:
  - Interest rate is really high for grade G borrowers
- Purpose vs interest Rate:
  - Small businesses end up defaulting more. We can see that the interest is also on the higher side for these borrowers.
- State Address vs Interest rate: Could not identify a pattern here. It was seemingly random.
- Funded amount vs interest rate:
  - Higher the funded amount, higher is the interest rate
- Interest Rate vs Term:
  - 5 year loans, the LC has decided that the rate of interest should be around 15% and for the 3 year loan, it is around 11%

(cont.)

- Funded amount vs Term:
  - larger funded amounts go with 5 year loan repayment plan. But there are so many outliers in the 3 year boxplot! Can't really say for sure.
- Annual income vs Grade:
  - Grade G borrowers are earning the higest salary while grade A borrowers have the lowest salary.
- Annual income vs Sub grade:
  - Sub-Grade x5(A5, B5, …, H5) borrowers are earning the higest salary while grade x1 borrowers have the lowest salary.(where x in set (A, B, C, D, E, F, G)

Conclusion

- We first cleaned the data to the best of our ability, exploiting redundancy and identifying patterns as we go.
- We then did *univariate* and *bivariate* analysis on the data and identified important driver variables that affected the loan status directly or indirectly.
- The final list of driver variables identified is present in the next slide.

Driver Variables identified

- Percentage of payment received(quite useful especially when used in conjunction with delta bins):
  - This information, however is not available at the time of loan application as earlier pointed out.
  - But if just look at this from a data analysis point of view, it does give good insights to predicting whether a currently running loan is going to be defaulted or not.
- Grade
- Purpose
- Annual Income:
- Funded amount
- Interest Rate
- Inquiries since last 6 months
- State Address
- Delta_bins(Time elapsed since loan issue date to the last payment date)

Recommendations

- Lower grade loans should be avoided as they have a higher chance of defaulting.
- We should try to lend money to people with a higher annual income and avoid ones on the lower side like 0 - 20,000 dollars.
- If the loan amount asked is on the higher side, there is a higher chance of defaults. Funded amounts in the range of 0 to 15,000 dollars have around the same chances of defaults. Loan amounts greater 15,000 dollars have a higher chance of defaulting. Therefore, it's best to stay within 0 - 15,000 dollars. May be within 20,000, if one is willing to take a little more risk. But safest option is within 15,000 dollars.
- Try to look out for loans with a lower interest rate, since they are less likely to be defaulted. Interest rates within 10% are the safest options. Anything beyond 14 % should definitely be avoided, and interest rates > 16% having the highest chances of defaulting.

Recommendations cont.

- Borrowers making more number of inquiries in the last 6 months are more likely to default. Borrowers who don't make any inquiries are the safest option. And borrowers making inquiries in the range 1-2 default more. Anything above that should definitely be avoided as the default rate >= 20%
- Borrowers from certain states tend to default more. Nebraska should definitely be avoided as 60% of the loans have been defaulted. Nevada(NV) is far second with around 21.73% chance of defaulting. So a table can be shared to lenders to make an informed decision.