EDA CASE ANALYSIS

G P SHIVA PRASAD (gpshivaprasad@gmail.com)

KARTHIK PATNAYKUNI (karthik-p@hotmail.com)

OUTLINE

- Problem Definition & Goals of Analysis
- Approach to Problem Solving
- Analysis & Findings: Application Data
- Observations
- Conclusions
- Future Scope for Analysis

PROBLEM DEFINITION & GOALS OF ANALYSIS

- Financial institutions earn by giving loans to clients. There are two ways in which the institutions may lose:
  - Giving loans to people who are likely to default
  - Not giving loans to people who are likely to pay back

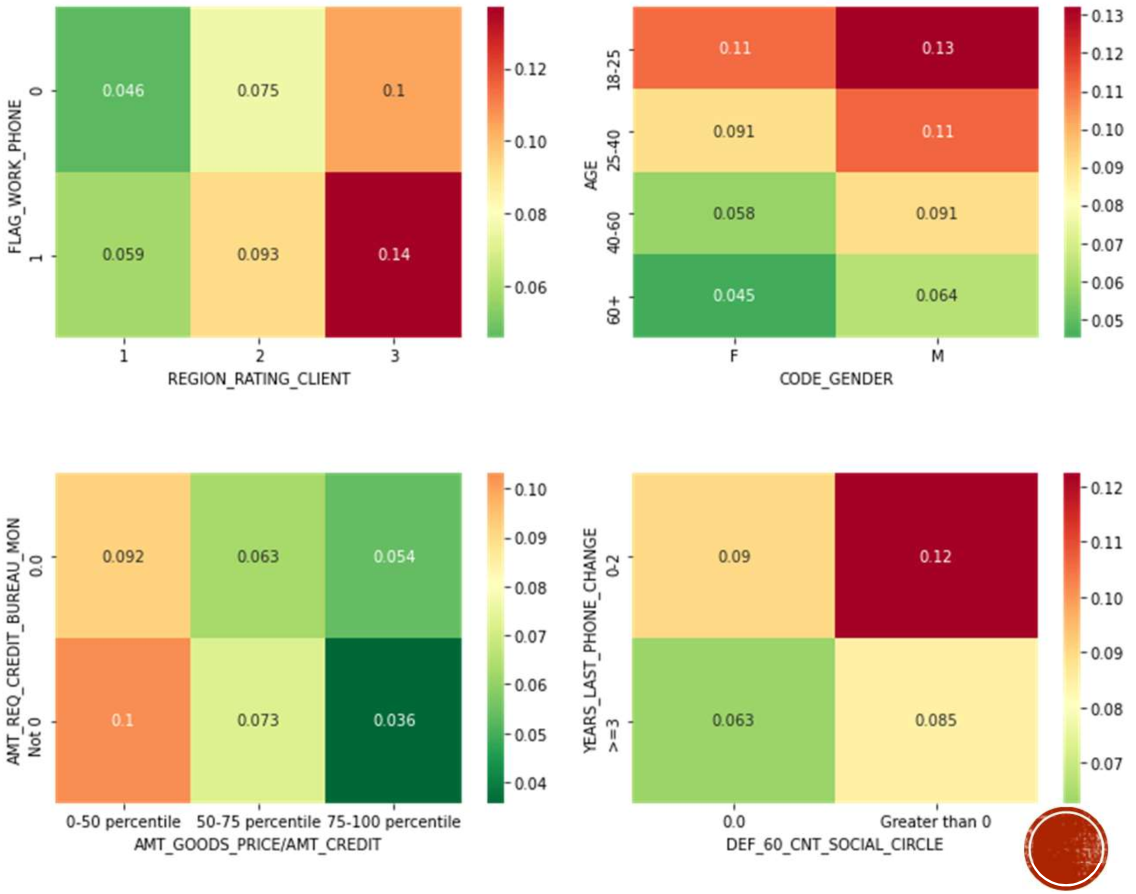
- Financial institutions can improve the earnings by following the virtuous cycle:
  - Capture as much data as possible about clients
  - Determine the data points that are strong predictors
  - Ensure that the most relevant data points are always captured
  - Give best interest rates to least likely to default and outperform competitors
  - Give higher interest rates to risky clients to reduce overall loss

APPROACH TO PROBLEM SOLVING

- General observation of the data in terms of no. of rows, columns, size of file, type of data etc.
- Deleting all columns where the missing values are huge (>40%)
- Column wise analysis for type of data it is, if there are outliers, how we intend to use each column in the analysis
- As the target column is Boolean, we will try and convert all data points to categorical values i.e. for continuous variables, we will bin them into categories
- By looking at imbalance of the target variable, we shall drop columns where imbalance is minimal i.e. the mean score of the target variable with respect to each category should be significant. We will take 10% deviation from the mean as significant deviation as a rule of thumb.
- Narrow down to 5-7 variables that have the maximum imbalance
- Perform bivariate analysis using the variables with maximum imbalance
- Arrive at 2-3 factors that have a maximum combined influence on the target column
- Merge the data with respect to the applicant ID
- Arrive at more factors that can be used to determine the credit worthiness of an applicant

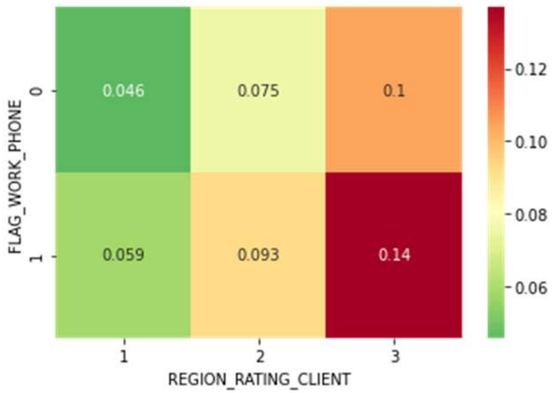
ANALYSIS & FINDINGS: APPLICATION DATA

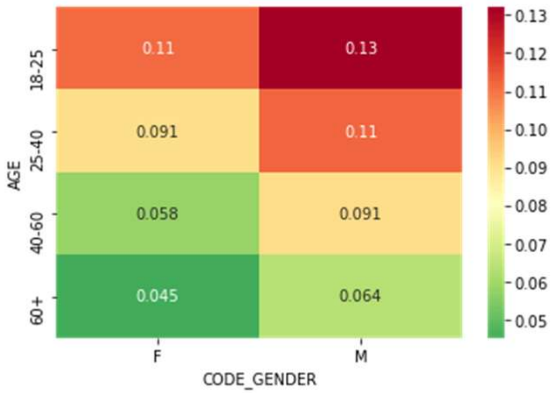
- Top Determinants of Credit Risk are:
  - Age
  - Gender
  - Flag_Work_Phone
  - Regions_Rating_Client
  - Years_Last_Phone_Change
  - Def_60_CNT_Social_Circle
  - Amt_Req_Credit_Bureau_Mon
  - Amt_Goods_Price/ Amt_Credit

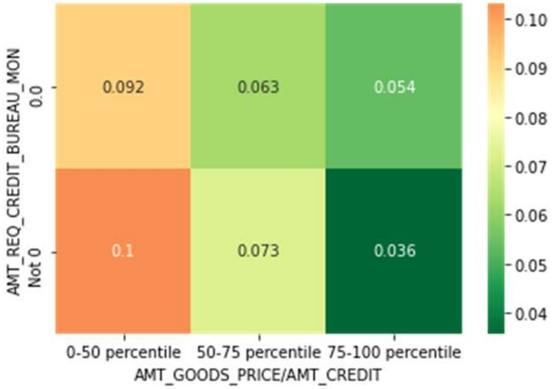


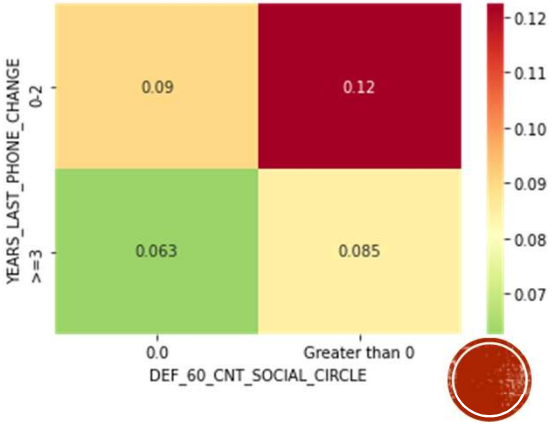
OBSERVATIONS

- Females as compared to males are more likely to repay
- With increasing age, the risk of default goes down
- People in region 3 are much more likely to default as compared to regions 1 and 2
- The ratio of the price of goods to be procured to credit is very good indicator of risk. The higher the value, the safer the loan is
- People who have not changed their phones for three or more years are less likely to default
- Default data from social circles is a good indicator of credit worthiness of an individual



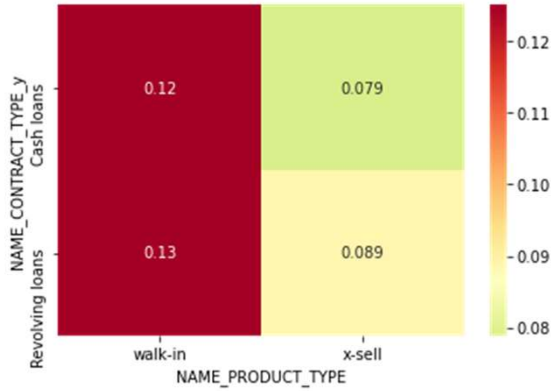


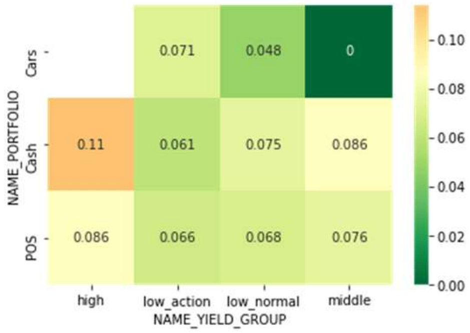


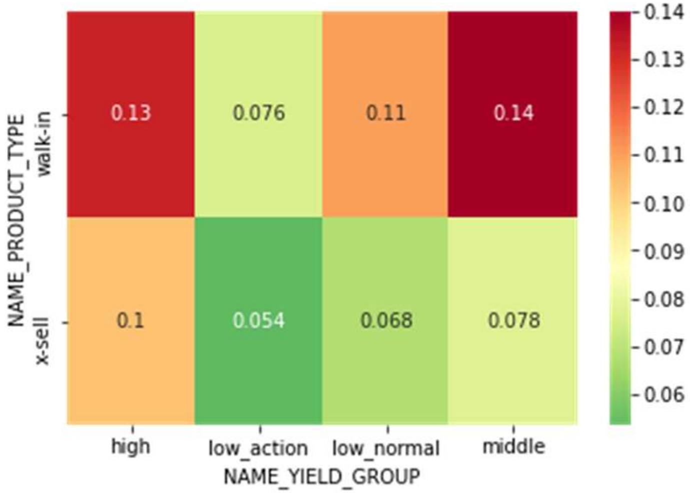


RECOMMENDATIONS

- Based on merged data the following recommendations can be given to a financial institutions, in ambiguous situations
  - Low_Action and Low_Normal yield groups are safer in general.
  - Walk-ins are riskier groups as compared to x-sell.
  - Cash transfers are consistently riskier as compared to POS, Cars.



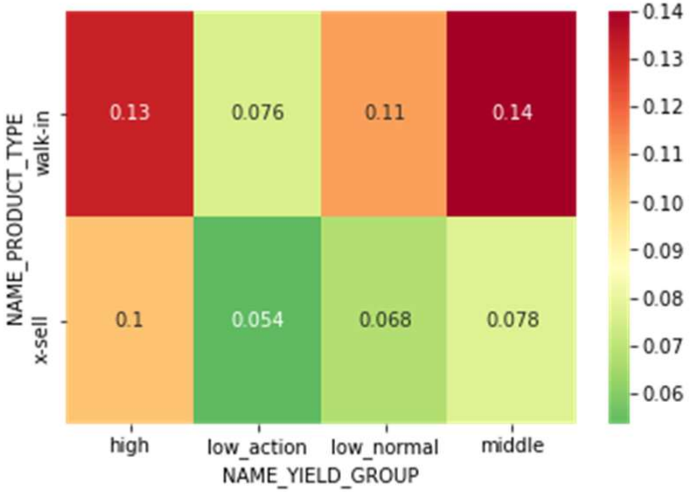




POSSIBLE CHANGES TO BE MADE IN THE BANK

- The high and middle yield groups are far riskier then the low action and low normal
- The banks historic data (Acceptance Rate of a Loan) is not in sync with this
- The bank can start improving the acceptance rate for low action and low normal yield groups and reduce acceptance on the high and middle groups

Name Yield Group	Count	Mean of Acceptance
XNA	1,56,111	0.5279
High	3,05,193	0.8486
Low_Action	77,766	0.7796
Low_Normal	2,70,241	0.7798
Middle	3,22,178	0.8483



FUTURE SCOPE FOR ANALYSIS

- While binning data, instead of using percentile scores, clustering using k-means or other techniques can deliver better result
- In-order to determine if a column is a strong influencer, a t-test can be performed rather than just taking an ad-hoc value like 10% imbalance.
- While deciding the credit worthiness of an individual micro clusters can be formed by taking several values in one go. Such micro-clusters can be formed by binning several categorical variables and placing the end results in each cluster.
- For riskier loans, one can start calculating a higher interest rate, so that the defaults are off set by higher returns.