CAPSTONE PROJECT

By

ARNAB MUKHERJEE

Business understanding and Data understanding are very critical first couple of steps for any data science project. Read the information given below and also refer to the data dictionary provided separately in an excel file to build your understanding.

Problem Statement:

A banking institution requires actionable insights from the perspective of Mortgage-Backed Securities, Geographic Business Investment andReal Estate Analysis.

The objective is to identify white spaces/potential business in the mortgage loan. The mortgage bank would like to identify potential monthly mortgage expenses for each of region based on factors which are primarily monthly family income in a region and rented value of the real estate. Some of the regions are growing rapidly and Competitor banks are selling mortgage loans to subprime customers at a lower interest rate. The bank is strategizing for better market penetration and targetingnew customers. A statistical model needs to be created to predict the potential demand in dollars amount of loan for each of the region in the USA. Also, there is a need to create a dashboard which would refresh periodically post data retrieval from the agencies. This would help to monitor the key metrics and trends.

The dashboard must demonstraterelationships and trends for the key metrics as follows: number of loans, average rental income, monthly mortgage and owner's cost, family income vs mortgage cost comparison across different regions. The metrics are described not to limit the dashboard to these few only.

Dataset Description

Following are the themes the fields fall under Home Owner Costs: Sum of utilities, property taxes.

- Second Mortgage: Households with a second mortgage statistics.
- Home Equity Loan: Households with a Home equity Loan statistics.
- Debt: Households with any type of debt statistics.
- Mortgage Costs: Statistics regarding mortgage payments, home equity loans, utilities and property taxes
- Home Owner Costs: Sum of utilities, property taxes statistics
- Gross Rent: Contract rent plus the estimated average monthly cost of utility features
- Gross Rent as Percent of Income Gross rent as the percent of income very interesting
- High school Graduation: High school graduation statistics.
- Population Demographics: Population demographic statistics.
- Age Demographics: Age demographic statistics.
- Household Income: Total income of people residing in the household.
- Family Income: Total income of people related to the householder.

Approach :

Following pointers will be helpful to structure your findings.

1. Import data

2. Figure out the primary key and look for the requirement of indexing

3. Gauge the fill rate of the variables and devise plans for missing value treatment.Please explainexplicitly the reason for the treatment chosen for each variable.

(Note: If you are using SAS, please ignore the portion of the question where visualization is required. Please write a macro for all data wrangling steps instead.)

4. Understanding homeowner costs are incredibly valuable because it is positivelycorrelated to consumer spending which drives the economy through disposable income. Perform debt analysis. You may want to follow the following steps:

   - Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10%.Visualize using geo-map. You may keep the upper limit for the percent of households with a second mortgage to roughly 50%.
   - Bad debt is the debt you should avoid at all costs such as a second mortgage or home equity loan. Conversely, Good debt is all other debt not including second mortgage or home equity loan.
     Bad Debt Equation:
     Bad Debt = P(Second Mortgage ∩ Home Equity Loan)
     Bad Debt = second_mortgage + home_equity - home_equity_second_mortgage
   - Create pie charts (Venn diagram) to show overall debt(% bad and good debt) and bad debt(2 mortgage and home equity loan).
   - Create Box and whisker plot and analyze the distribution for 2nd mortgage, home equity, good debt and bad debt for different cities.
   - Create a collated income distribution chart for family income, house hold income and remaining income.

5. PerformEDA and come out with insights intopopulation densityand age. You may require deriving new fields(Make sure to weight averages for accurate measurements):

   - Population density(hint-use 'pop' and 'Aland' to calculate)
   - median age (hint-use the variables 'male_age_median', 'female_age_median','male_pop', 'female_pop')

   Visualize the findings using appropriate chart type.

6. Create bins for populationinto a new variableby selecting appropriate class interval so that the no of categories(bins) don't exceed 5 for the ease of analysis. Analyze the married, separated and divorced population for these population brackets. Visualize using appropriate chart type.

7. Please detail your observations for rent as a percentage of income at an overall level and for different states.

8. Perform correlation analysis for all the relevant variables by creating a heatmap.Describe your findings.

9. The economic multivariate data has a significant number of measured variables. The goal is to find where the measured variables depend on a number of smaller unobserved common factors or latent variables. Each variable is assumed to depend on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as "specific variance" because it is specific to one variable.Obtain the common factors

and then plot the loadings.Use factor analysis to find latent variables in our dataset and gain insight into the linear relationships in the data

→ Highschool graduation rates
→ Median population age
→ Second Mortgage Statistics
→ Percent Own
→ Bad Debt Expense

10. Build a linear Regression model to predict the total monthly expenditure for home mortgages loan; please refer - 'deplotment_RE.xlsx'.
    Column hc_mortgage_mean is predicted variable. This is mean monthly mortgage and owner costs of specified geographical location.
    Note: *Exclude loans from prediction model which have NaN values for hc_mortgage_mean. NaN represents not a number/missing values.*

    - Run a model at a Nation level. If the accuracy levels and R square are not satisfactory proceed to below step
    - Run another model at State level. There are 52 states in USA.

    Considerations: Keep below considerations while building a linear regression model

    - Variables should have significant impact on predicting Monthly mortgage and owner costs
    - Utilize all predictor variable to start with initial hypothesis
    - R square of 60% and above should be achieved
    - Ensure Multi-collinearity does not exist in dependent variables
    - Test if predicted variable is normally distributed

11. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business.The dashboard must entail the following:

    a) Box plot of distribution of average rent by type of place (Village, urban, town etc.)
    b) Pie charts (Venn diagram) to show overall debt (% bad and good debt) and bad debt (2 mortgage and home equity loan)
    c) Explore the top 2,500 locations where the percentage of households with a second mortgage is the highest and percent ownership is above 10%. Visualize using geo-map.
    d) Heat map for correlation matrix
    e) Pie chart to show the population distribution across different types of places (Village, urban, town etc.)