Data due Diligence Project – Week 5

Merrimack College

Data Exploration

Amol Gote

(Data science student)

June 16th, 2020

Feature engineering: For the given data set there are 8 columns which are with missing values, some of the data attributes are of lesser significance for the data analysis and business objective, these include Pets related variables so missing values for these variables have been ignored. Remaining 4 variables include HouseholdSize, HomeOwner, Gender, JobCategory, these variables would help in understanding the customer base better and could be used for data segmentation, so applied appropriate imputation techniques on it. Gender data distribution is normal and is almost equal, so applied equal distribution of male and female for 33 missing values, this way variance and distribution of the variable are preserved. For HouseholdSize mean and median are almost the same that indicates normal data distribution, so applied median imputation with the value of 2. For homeownership, data is not evenly distributed, customers with homeownership are considerably more, so applied mode imputation technique to it. For JobCategory created a separate “Misc” category for missing values. For all the currency variables removed dashes, $ signs, empty spaces, and replaced it with NA, so that statistical computation can be performed and data is consistent across all variables. Commute time is normally distributed so applied mean imputation for 2 rows with missing values.

For statistical analysis added following variables TotalDebt, LastMonthTotalValue, OverTenureTotalValue, AverageMonthlyRevenue. These variables could help in finding high-value customers and if any relationship exists between debt and total value that the customer pays. Finally trimmed down the dataset to limit the number of variables of interest these include some from customer profile and some from customer relationship values as well as additional computed variables.

Here are the key findings from the dataset based on statistical analysis gender does not have an impact on the median ($22-23) or the average monthly revenue ($35-36) from customers. There is no age bias and the frequency of customers is nearly equal across the age distribution for 18 to 70 years of age. The average monthly revenue generated across the various age segments (buckets of 10) is the same around $30k, with a minimum of $24k for the age group 70-80 as there are a smaller number of customers. Maximum monthly revenue is generated from customers with total debt of less than 5% and is almost greater than $100k, 5-10% debt is $38K and 10-15% is $16K. Beyond 15% of debt, total revenue generated keeps dropping. For a household size of 1 maximum monthly revenue of $69K is generated, followed by a household size of 2. There are 1481 customers with household size greater than 1 but are not multiline, these could be potential target customers for group deals or family plans. College-educated students generate the highest monthly revenue with education years from 15 to 17 years. Customers with the household income of 20-30K are the ones which generate maximum monthly revenue, followed by 30-40K, and then monthly revenue tapers of as household size increases. This presents potential business opportunities to target customers with high household income. The job category of “Sales” and “Professional” are the ones that generate maximum monthly revenue close to $50K. Customers with homeownership ($120K) generate double the monthly revenue than that of non-homeowners ($60k). Customers with a debt to income (DTI) ratio of 0-10 are the ones that generate maximum monthly revenue ($100K), followed by 10-20%. Commute time has no linear relationship with data revenue, in general customers with 20-30 minutes commute time consume highest data and generate maximum monthly revenue as well.

In general customers with DTI of 0-20%, with homeownership, who have a household income of $20-50K, with a college degree, with a household size of 1 or 2, with total debt less than 5%, with job category of “Sales” or “Profession” are high-value customers for the telecommunications company.