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 NumberPets, NumberCats, NumberDogs, NumberBirds so missing values for these variables has 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 same that indicates normal data distribution, so applied median imputation with value of 2. For HomeOwner, data is not evenly distributed customers with home ownership are considerably more, so applied mode imputation technique to it. For JobCategory created a separate “Misc” category. 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. CarOwnership, CarBrand, CarValue are related to each other so have replaced -1 as missing value (NA) and CarValue as 0 and for treating these variables either a prediction model or KNN Imputation can be used. CommuteTime is normally distributed so applied mean imputation for 2 rows with missing values, commute time could be related to data consumed, as customers while commuting tend to use mobile device more for either entertainment or reading.

For statistical analysis added following variables TotalDebt = CreditDebt + OtherDebt LastMonthTotalValue = VoiceLastMonth + EquipmentLastMonth + DataLastMonth, OverTenureTotalValue = VoiceOverTenure + EquipmentOverTenure + DataOverTenure. These variables could help in finding high value customers and if any relationship exists between debt and total value that customer pays.