A picture containing knife

Description automatically generated

Module 2 Assignment

**Generalized Linear Models**

ALY 6020 – Predictive Analysis

Instructor – Justin Grosz

Name -Ashish Kothari NUID- 001360980

**Introduction**

**Generalized Linear models:**

Generalized Linear Models are widely used in predicting or statistics problems. This includes various models like ANNOVA, linear regression and so on. This also tell us how much the variables are important and also using this we can predict if a variable is important or not.

In this assignment we are going to check different models and compare the best model fit with the KNN model. We have marketing\_campaign dataset which includes 2240 rows and 29 variables. It has both categorical and integer values. In this we are trying to predict what drives the customer to subscribe for a magazine. In this way we can help the marketing campaign to target audience accordingly.

Steps:

1. Check the data for missing values and outliers
2. Choose the right model
3. Build one model
4. Compare the model with KNN model
5. Provide suggestions on what can be done for better marketing.

**Step 1:**

**Checking the dataset**

As we see that there are total 2240 rows and 29 variables. We have some variables as factor and some as integer. We need to change the class of factor variables to character variable.

We observe that the three variables were factor

1. Education (factor -> character)
2. Marital\_Status (factor -> character)
3. Dt\_customer (factor -> numeric) – We have also used standardized date format and using ymd function in R.

**Checking for outliers**

As we check for outliers, we use boxplot function and check which variables has most outliers.

After boxplot we notice that Income variable has outliers and that has to be treated. Also, year\_birth has outlier but that are supposed to be ignored as it has wide range of people born in different year.

After this we check the outlier’s value and based on that we can decide which are to be removed. We observed that Income has 8 outliers and that can be treated.

> OutVals [1] 157243 162397 153924 160803 157733 157146 156924 666666

A picture containing graphical user interface

Description automatically generated

**Treating outliers:**

After checking the values, we decide to remove the outliers for a better accuracy and prediction. Using %in%outliers we remove the outliers.

**Checking for missing values:**

We notice that Income variable has missing values. We check by using summarize function. We use colSums function to calculate number of NA values. Now we calculate average to check if the NA can be replaced by average or not. We calculate average by dropping the NA values and also by replacing NA= average value of Income.

We notice as number of NA value is not much so we can replace the NA values by average of Income values. We use replace\_na function to replace the missing value.

**Step 2:**

We need to check the relation between variables. For this we need to build a correlation matrix. But here we have an issue that we have two character variables which we cannot use for correlation matrix.

We check the levels for both the variables Education and Marital\_Status. We replace the labels with levels, and we can convert these variables into numeric.

For Education - levels = c("2n Cycle","Basic","Graduation","Master","PhD"),

labels = c(1,2,3,4,5)))

For Marital\_Status-

levels = c("Absurd","Alone","Divorced","Married","Single","Together","Widow","YOLO"),

labels = c(1,2,3,4,5,6,7,8)))

Now we convert these two variables into numeric.

**Correlation Matrix:**

Now as all the variables are numeric and there are no categorical variables, we create correlation matrix and check for the variables that have highest relation with Response variable. We select our best 5-6 variables to further move with the analysis. The best six variables are MntWines, MntMeatProducts, NumCatalogPurchases, AcceptedCmp3, AcceptedCmp5, AcceptedCmp1

Table

Description automatically generated

We use GLM and linear regression models as we have all the numerical values and we can easily predict if the variables are selected properly or not. Also, it is easy to look and predict if the model is a good fit or not. Therefore, we use these two and we will compare the results with KNN model.

**Step 3:**

We build GLM model and Linear Regression model for our prediction for the magazine.

We build Linear regression model using lm function and we use summary function to check for the variables and accuracy of the model. We see that R2 value is very low and the model is only 19% accurate. Therefore, this model is not a good fit.

**GLM model:**

In this model we use different functions and also create a confusion matrix. We use debt.glm and glm function to create this model. We get RMSE value use RMSE function and by using test dataset with predicting variable “Response”.

We get the RMSE value and also, we get the accuracy of the model. We see this model is 34% fit which is better than the linear regression model. Also, it shows good relation with variables. So, we can include those variables and take into consideration.

**Step 4:**

**KNN model:**

We have already built KNN model last assignment, so we use the similar concepts to build a KNN model for this dataset. We create KNN model for k=1 to k=20 and see that graph is not a complete dip. It is forming a U shaped so we can predict that is it a good fit. KNN has 41.9% for k=1, 37.6% for k=2 and it gradually reduces as the value of k increases. So we can conclude that accuracy is more than compared to above all the models. We can say that so far KNN model is the best fit model for this dataset.

Chart, line chart

Description automatically generated

**Step 5:**

**Suggestions or Feedbacks:**

After cleaning and analyzing the dataset of marketing campaign, we can say that Stakeholders should concentrate on people buying Meat products and Wines. This people are the ones who are more likely to subscribe to the magazine. Also, stakeholders should keep contacting the people who subscribed their magazines during Campaign 1, 3 and 5. To increase and bring in new and more numbers of subscribes the target audience would be the one who buys meats and wines.