**PART 1- Linear Regression & Cross Validation**

**Q1) Use set.seed(8). Divide the data set into training set (80% of the data) and test\_set (20%)**

**Ans.** After fetching the dataset, it was divided into two separate sets naming training and test data set after setting the seed to 8 as asked in the question. The training data set which is comprised of 80% of the dataset observation where as the test data set is the remaining 20% of the observations. The original dataset has 12 variables and more than five hundred thousand of observations. The training data set has 429978 observations, whereas teat data set has 107433 observations.

**Q2) Build a linear regression model to predict Purchase using predictors Gender, Age, Occupation, Stay\_In\_Current\_City\_Years, Marital\_Status, Product\_Category\_1, City\_Category. Do not use other variables and make sure categorical variables are considered as “factor” before you build the model.**

**Ans.** The linear regression model is made from given predictor variables Gender, Age, Occupation, City\_Category, Stay\_In\_Current\_City\_Years, Marital\_Status and Product\_Category\_1. Out of which the categorical variables were specifically considered as factors and validated via code. Only two variables naming occupation and Product\_Category\_1 were not considered as factors for building the regression model. Below are the results of the regression model build for training and test data sets separately.

**Residuals:**

Min 1Q Median 3Q Max

-10149.9 -3143.5 -666.2 2288.5 17642.7

**Summary**

Residual standard error: 4709 on 430044 degrees of freedom

Multiple R-squared: 0.1069, Adjusted R-squared: 0.1069

F-statistic: 3217 on 16 and 430044 DF, p-value: < 2.2e-16

**Conclusion:**

The slope value is not zero. As the F-statistic is considerably large and p-value is less than 0.05 which provides enough evidence to reject the null hypothesis so we can claim the results are statistically significant.

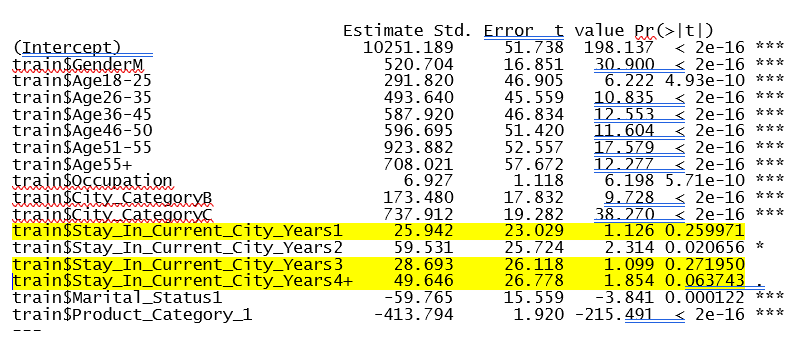
**Q3) Which variables are significant predictors of Purchase?**

**Ans.** The Stay\_In\_Current\_City\_Years variable is significant predictor of Purchase for various levels are listed below.

Stay\_In\_Current\_City\_Years1

Stay\_In\_Current\_City\_Years3

Stay\_In\_Current\_City\_Years4+



**Q4) Predict Purchase amount using the model you built separately for training set and test set. Calculate the Square root of Mean Square Errors for training and test set predictions. Test set RMSE (root mean squared error) will be the holdout error value.**

**Ans.**

|  |  |  |
| --- | --- | --- |
| **Value Type** | **Training** | **Test** |
| RMSE | 4709.37 | 10468.41 |
| Holdout Error Value | N/A | 10468.41 |

**Q5) Using cv.glm function, perform a 5-fold cross validation on Black Friday data set using the same predictors and response indicated above in Question2. Calculate cross validation RMSE and compare to the value you found in Q4 using holdout method. Which method would you prefer?**

**Ans.**

**PART 2: Logistic Regression**

**Use set.seed (9) and only holdout method for this analysis -80:20 ratio again for train and test sets**

**Q6) Create a histogram of Product\_Category\_1.**

**Ans. A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

**Q7) Create a table that shows the counts(frequency) of each Product\_Category\_1 using table() function). How many different categories were observed in this variable in BlackFriday data set?**

**Ans.** Eighteen different categories were observed in this variable in BlackFriday data set. Shown below is the table for various categories in this variable.

|  |  |
| --- | --- |
| **Product\_Category\_1** | **Frequency** |
| 1 | 138353 |
| 2 | 23499 |
| 3 | 19849 |
| 4 | 11567 |
| 5 | 148592 |
| 6 | 20164 |
| 7 | 3668 |
| 8 | 112132 |
| 9 | 404 |
| 10 | 5032 |
| 11 | 23960 |
| 12 | 3875 |
| 13 | 5440 |
| 14 | 1500 |
| 15 | 6203 |
| 16 | 9697 |
| 17 | 567 |
| 18 | 3075 |

**Q8) Using the output from Q7 above, calculate the probability of occurrence of each level in Product\_category\_1.**

**Ans.** The probabilities of occurrence of each level of Product\_category\_1 are listed below.

X.prod.Freq..sum.prod.Freq.

1 0.2573640613

2 0.0437128077

3 0.0369230826

4 0.0215169176

5 0.2764106351

6 0.0375090452

7 0.0068232086

8 0.2085877930

9 0.0007515202

10 0.0093605195

11 0.0445703592

12 0.0072082697

13 0.0101194806

14 0.0027902979

15 0.0115388121

16 0.0180383461

17 0.0010547326

18 0.0057201108

**Q9) Determine the levels of Product\_category\_1 which has lower than 0.03 probability of occurrence and consider these levels as “low probability levels”.**

**Ans.** The levels of Product\_category\_1 which has lower than 0.03 probability of occurrence and which can be considered as “low probability levels” are listed as under.

X.prod.Freq..sum.prod.Freq.

[4,] "0"

[7,] "0"

[9,] "0"

[10,] "0"

[12,] "0"

[13,] "0"

[14,] "0"

[15,] "0"

[16,] "0"

[17,] "0"

[18,] "0"

**Q10) Create a new variable that takes the value of 0 if an observation’s Product\_category\_1 level belongs to the “low probability group” described above in Q9. If the level does not belong to the low probability group, new variable will take the value of 1. Add this newly created binary variable to BlackFriday data set.**

**Ans.**

**Q11) Build a logistic regression model that predicts whether a purchase will belong to a low probability level or not. Use predictors Gender ,Age, Occupation, City\_Category, Stay\_In\_Current\_City\_Years, Marital\_Status,Purchase.**

**Ans.**

**Q12) Calculate Test set misclassification rate and accuracy of your model using different threshold(cutoff) probabilities. Which cutoff value would you choose? Why?**

**Ans.**