**ISE 240**

**Report**

**On**

**Black-Friday Sales**

**[](https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjEotKp1vfdAhXmHDQIHXm7CP8QjRx6BAgBEAU&url=https://techcrunch.com/2013/07/19/san-jose-states-bold-experiment-in-online-ed-disappoints-suspends-pilot-with-udacity/&psig=AOvVaw1PiEQrTriKUo8qxKXlAOWu&ust=1539116601226785)**

**Submitted To**

**S. Ayca Erdogan**

**on**

**March 29, 2019**

**Group:**

**Shikhar Chawla**

**Vivek Shah**

**Siddhanth Vashisht**

1. **Executive Summary**

An analysis was conducted on the data of Black-Friday Sales (sample of transactions made in a retail store). The purpose of this data analysis was to understand the customer purchase behavior against different products. The data contained 12 different types of variables, out of which some of the variables were categorical. Purchase was considered as a dependent variable.

Multiple Linear Regression was done on the dataset which fetched an R-sq. value of 63.18%.

A Logistic Regression Model was also built for the given dataset which helped us achieve the accuracy of 71.32613%.

1. **Methodology**
2. **Part 1**

The data was divided into 2 sets for training and testing, and a multiple regression model was built. Following this, significant predictors of Purchase were identified. After Identifying the significant predictors, prediction for Purchase amount for training set and the test set was done and their RMSE (root mean square error) was calculated. The last step for part one was to perform a K-fold cross validation on the Black-Friday dataset and obtain RMSE for it, leading to comparing RSME’s of prediction and K-fold cross validation.

1. Dividing the Black-Friday Dataset

The dataset was divided into separate parts, 80% of the data was used to train the model and the remaining 20% of the data was used to test the model. For dividing the data, seed was set at 8 and the “sample” function was used.

1. Multiple Linear Regression Model

A multiple linear regression model was built using the training set, for this step the “lm” function was used. Prior to this step some of the variables which were not categorical were converted into categorical variables and this was done by using the “as.factor” function.

1. Identifying Significant Predictors

Significant predictors were identified based on the P-values obtained from the regression output. Gender, Product category\_1, Marital\_Status, City\_Category and Occupation were identified as significant predictors of purchase because their P-values were less than alpha(0.05).

1. Predicting Purchase Amount

The Purchase amount was predicted for the training set and the test set along with their RMSE’s. This was done using the “predict” function.

1. K-Fold Cross Validation

K-Fold Cross Validation was performed using the “cv.glm” function and the value of K was set at 5, following this the RSME for the same was calculated. RMSE for K-Fold CV and Prediction were compared.

1. **Part B**

A logistic regression model was built in a similar way by dividing the data into test set and training set. Selective predictor variables were used as mentioned. Histogram plot was created for the predictor variable Product Category 1 to see the distribution of data. The categorical variable Product Category 1 was converted to binary variable by dividing it on the basis of the occurrence of each category. Logistic Regression analysis was then conducted on the training set and was used on the test set to get the miscalculation rate.

6) Plotting Product Category 1

Using the “hist()” the distribution of the variable Product Category 1 was found out.(output in appendix A.2).

7) Table for the variable

Using the “table()” function the frequency of each of the 18 categories of Product Category 1 was found out. (Output in appendix A.3).

8) Converting Product Category 1 into a binary variable

The predetermined probability of 0.03 was used to convert the variable into binary variable. Each of the Product Category 1 whose probability was less than 0.03 was named “low probability level” and then numbered 0. All the other levels of Product Category 1 whose probability was greater than 0.03 was named “high probability level” and then numbered 1. This was done using an if else function inside a for loop.

9) Logistic Regression Model

A logistic regression model was carried out on the training set between the Probability level of Product Category 1 and the predictor variables that were Gender, Age, Occupation, Purchase, City\_Category, Stay\_In\_Current\_City\_Years, Marital\_Status. This model was then used to predict the responses in the test set. Different threshold values were tested, and that value of threshold was selected which reduced the misclassification rate and predicted both categories of 0 and 1 reasonably was selected.

1. **Results & Conclusion**
2. Multiple Linear Regression

R-sq. = 63.18%

Adj. R-sq. = 63.17%

Conclusion – High F value and low p value suggests that the model is significant and at least one of the slopes of multiple regression model is not equal to zero, ( output in appendix A.1).

1. Prediction (MSE, RMSE)

MSE Training Set = 9144370

MSE Test Set = 9178468

RMSE Training Set = 3023.966

RMSE Test Set = 3029.599

1. K-Fold Cross Validation

MSE K-Fold Cross Validation = 9152878

RMSE K-Fold Cross Validation = 3025.372

Conclusion - Hold out error value and cross validation RMSE are very close to each other, but cross validation RMSE is smaller than holdout error value, hence 5-fold cross validation should be used.

1. Logistic Regression

Misclassification Rate: 28.67387%

Accuracy: 71.32613%

False positive rate = 20.87%

False negative rate= 7.8%

Conclusion – The model is predicting approximately 71.3% accurately when tested on test data, (output in appendix A.4).

1. Threshold

Conclusion - The threshold of 0.7 was selected because the model predicted very high probability for every case, therefore, different threshold values were tested, and that value was selected which reduced false negative rate and predicted both categories of 0 and 1 reasonably.

1. **Appendix**

A.1) Following is the result from multiple linear regression done in R:

Call:

lm(formula = Purchase ~ Gender + Age + Occupation + City\_Category +

Stay\_In\_Current\_City\_Years + Marital\_Status + Product\_Category\_1,

data = Black\_friday\_data, subset = train)

Residuals:

Min 1Q Median 3Q Max

-15540.5 -1611.8 413.2 1966.2 8410.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13368.359 49.507 270.027 < 2e-16 \*\*\*

GenderM -47.756 11.229 -4.253 2.11e-05 \*\*\*

Age18-25 -197.777 46.450 -4.258 2.06e-05 \*\*\*

Age26-35 -92.060 46.322 -1.987 0.046878 \*

Age36-45 31.925 46.984 0.679 0.496824

Age46-50 58.495 49.064 1.192 0.233181

Age51-55 302.722 49.617 6.101 1.05e-09 \*\*\*

Age55+ 190.975 52.571 3.633 0.000280 \*\*\*

Occupation1 -18.988 20.619 -0.921 0.357089

Occupation2 80.345 24.780 3.242 0.001186 \*\*

Occupation3 236.074 28.886 8.173 3.03e-16 \*\*\*

Occupation4 174.203 19.704 8.841 < 2e-16 \*\*\*

Occupation5 53.106 33.486 1.586 0.112757

Occupation6 216.465 27.607 7.841 4.48e-15 \*\*\*

Occupation7 128.768 19.335 6.660 2.74e-11 \*\*\*

Occupation8 -244.985 87.977 -2.785 0.005359 \*\*

Occupation9 126.302 45.920 2.750 0.005951 \*\*

Occupation10 -82.759 49.744 -1.664 0.096174 .

Occupation11 138.023 34.435 4.008 6.12e-05 \*\*\*

Occupation12 301.645 23.434 12.872 < 2e-16 \*\*\*

Occupation13 128.161 44.539 2.877 0.004009 \*\*

Occupation14 205.484 24.500 8.387 < 2e-16 \*\*\*

Occupation15 388.289 33.754 11.504 < 2e-16 \*\*\*

Occupation16 116.430 25.313 4.600 4.23e-06 \*\*\*

Occupation17 191.046 21.684 8.811 < 2e-16 \*\*\*

Occupation18 3.862 44.151 0.087 0.930290

Occupation19 -285.367 39.589 -7.208 5.68e-13 \*\*\*

Occupation20 -82.417 22.827 -3.610 0.000306 \*\*\*

City\_CategoryB 127.189 11.524 11.037 < 2e-16 \*\*\*

City\_CategoryC 552.736 12.499 44.221 < 2e-16 \*\*\*

Stay\_In\_Current\_City\_Years1 -3.971 14.846 -0.267 0.789097

Stay\_In\_Current\_City\_Years2 35.032 16.563 2.115 0.034427 \*

Stay\_In\_Current\_City\_Years3 -18.776 16.849 -1.114 0.265122

Stay\_In\_Current\_City\_Years4+ 18.608 17.255 1.078 0.280852

Marital\_Status1 -58.271 10.031 -5.809 6.29e-09 \*\*\*

Product\_Category\_12 -2360.379 23.874 -98.867 < 2e-16 \*\*\*

Product\_Category\_13 -3514.189 25.806 -136.177 < 2e-16 \*\*\*

Product\_Category\_14 -11266.466 32.644 -345.126 < 2e-16 \*\*\*

Product\_Category\_15 -7361.802 12.705 -579.456 < 2e-16 \*\*\*

Product\_Category\_16 2234.639 25.527 87.540 < 2e-16 \*\*\*

Product\_Category\_17 2819.221 56.492 49.905 < 2e-16 \*\*\*

Product\_Category\_18 -6116.035 13.701 -446.384 < 2e-16 \*\*\*

Product\_Category\_19 1868.172 166.731 11.205 < 2e-16 \*\*\*

Product\_Category\_110 6016.330 48.932 122.953 < 2e-16 \*\*\*

Product\_Category\_111 -8908.403 23.657 -376.567 < 2e-16 \*\*\*

Product\_Category\_112 -12306.370 54.938 -224.007 < 2e-16 \*\*\*

Product\_Category\_113 -12898.550 46.811 -275.544 < 2e-16 \*\*\*

Product\_Category\_114 -427.452 88.229 -4.845 1.27e-06 \*\*\*

Product\_Category\_115 1157.797 44.078 26.267 < 2e-16 \*\*\*

Product\_Category\_116 1133.031 35.404 32.003 < 2e-16 \*\*\*

Product\_Category\_117 -3588.793 143.227 -25.057 < 2e-16 \*\*\*

Product\_Category\_118 -10683.241 61.747 -173.016 < 2e-16 \*\*\*

---

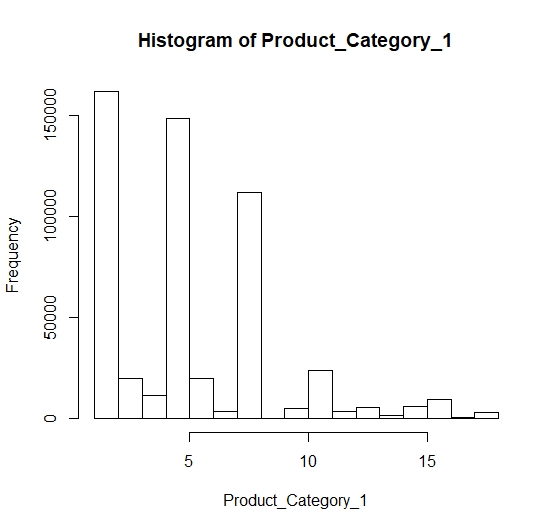
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3024 on 430010 degrees of freedom

Multiple R-squared: 0.6318, Adjusted R-squared: 0.6317

F-statistic: 1.447e+04 on 51 and 430010 DF, p-value: < 2.2e-16

A.2) Following is the histogram of product category 1:



A.3) Following is the frequency table of each level of product category 1:

|  |  |
| --- | --- |
| Levels | 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 |

A.4) Following is the result from logistic regression done in R:

|  |
| --- |
| Call: |
| glm(formula = Probability\_level ~ Gender + Age + Occupation + |
| Purchase + City\_Category + Stay\_In\_Current\_City\_Years + Marital\_Status, |
| family = "binomial", data = Black\_friday\_data[train\_logistic, |
| ]) |
|  |
| Deviance Residuals: |
| Min 1Q Median 3Q Max |
| -1.8481 -1.4702 0.7397 0.7808 1.0331 |
|  |
| Coefficients: |
| Estimate Std. Error z value Pr(>|z|) |
| (Intercept) 1.091e+00 3.604e-02 30.273 < 2e-16 \*\*\* |
| GenderM -8.871e-03 8.497e-03 -1.044 0.296471 |
| Age18-25 1.721e-01 3.387e-02 5.082 3.74e-07 \*\*\* |
| Age26-35 2.264e-01 3.380e-02 6.699 2.10e-11 \*\*\* |
| Age36-45 1.426e-01 3.429e-02 4.159 3.20e-05 \*\*\* |
| Age46-50 1.103e-01 3.589e-02 3.075 0.002107 \*\* |
| Age51-55 1.697e-01 3.638e-02 4.666 3.07e-06 \*\*\* |
| Age55+ 2.078e-01 3.881e-02 5.356 8.52e-08 \*\*\* |
| Occupation1 1.166e-01 1.572e-02 7.416 1.21e-13 \*\*\* |
| Occupation2 5.725e-02 1.876e-02 3.052 0.002275 \*\* |
| Occupation3 6.586e-02 2.196e-02 3.000 0.002703 \*\* |
| Occupation4 5.759e-02 1.487e-02 3.873 0.000107 \*\*\* |
| Occupation5 -7.860e-03 2.511e-02 -0.313 0.754301 |
| Occupation6 7.393e-02 2.092e-02 3.534 0.000409 \*\*\* |
| Occupation7 1.246e-01 1.472e-02 8.462 < 2e-16 \*\*\* |
| Occupation8 6.175e-02 6.677e-02 0.925 0.355053 |
| Occupation9 -1.289e-02 3.415e-02 -0.377 0.705951 |
| Occupation10 -1.432e-01 3.607e-02 -3.971 7.15e-05 \*\*\* |
| Occupation11 9.824e-02 2.642e-02 3.719 0.000200 \*\*\* |
| Occupation12 6.018e-02 1.772e-02 3.396 0.000683 \*\*\* |
| Occupation13 9.983e-02 3.401e-02 2.936 0.003329 \*\* |
| Occupation14 1.439e-01 1.879e-02 7.656 1.91e-14 \*\*\* |
| Occupation15 1.001e-01 2.582e-02 3.875 0.000107 \*\*\* |
| Occupation16 4.496e-02 1.911e-02 2.353 0.018637 \* |
| Occupation17 4.242e-02 1.630e-02 2.602 0.009276 \*\* |
| Occupation18 -1.357e-01 3.220e-02 -4.213 2.52e-05 \*\*\* |
| Occupation19 -4.244e-02 2.935e-02 -1.446 0.148188 |
| Occupation20 -1.598e-02 1.708e-02 -0.936 0.349497 |
| Purchase -2.953e-05 6.936e-07 -42.569 < 2e-16 \*\*\* |
| City\_CategoryB -1.449e-02 8.736e-03 -1.659 0.097126 . |
| City\_CategoryC 4.547e-02 9.520e-03 4.776 1.79e-06 \*\*\* |
| Stay\_In\_Current\_City\_Years1 2.668e-02 1.120e-02 2.382 0.017237 \* |
| Stay\_In\_Current\_City\_Years2 2.294e-02 1.251e-02 1.833 0.066731 . |
| Stay\_In\_Current\_City\_Years3 3.118e-02 1.273e-02 2.449 0.014320 \* |
| Stay\_In\_Current\_City\_Years4+ 1.196e-02 1.302e-02 0.919 0.358319 |
| Marital\_Status1 -1.480e-02 7.619e-03 -1.942 0.052096 . |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| (Dispersion parameter for binomial family taken to be 1) |
|  |
| Null deviance: 490971 on 430060 degrees of freedom |
| Residual deviance: 488532 on 430025 degrees of freedom |
| AIC: 488604 |
|  |
| Number of Fisher Scoring iterations: 4 |