PREDICTING THE PURCHASE DONE ON BLACK FRIDAY

RAHUL KUMAR PATRO*
IMSC MATHEMATICS AND COMPUTING
BIRLA INSTITUTE OF TECHNOLOGY,MESRA
RANCHI , 835215
rahulkpatro@gmail.com

SUDAKSHINA BHATTACHARJEE
IMSC MATHEMATICS AND COMPUTING
BIRLA INSTITUTE OF TECHNOLOGY,MESRA
RANCHI , 835215
sudakshinabit27@gmail.com

ANKIT TEWARI
Artificial Intelligence Engineer
Knowledge Engineering and Machine Learning Group
ankit.tewari@estudiant.upc.edu

July 6, 2019

ABSTRACT

Making predictions is something a business or system depends on, which is indirectly dependant on data analytics. Data analytics is the science of analyzing raw data in order to make conclusions about that information.This analysis can then be used to optimize processes to increase the overall efficiency of a business or system. The dataset here, that we have obtained from Kaggle, is a sample of the transactions made in a retail store on Black Friday. The store wants to know better the customer purchase behaviour against different products.

***K*eywords** Multiple Linear Regression · KNN Regression · Data Analytics

1 INTRODUCTION–

This project aims to develop a sales prediction model using methods like Linear Regression and K Nearest Neighbours to complete this project. The system will process the data.
Processing of the data has several sub divisions like reading of the data, splitting the dataset for performing tasks like training and testing to perform the required prediction.

The various tasks involved are 1)Reading data from the CSV file 2) Processing of data to make it usable for the required prediction model 3)Training model to find out the prediction with the help of Multiple Linear Regression and KNN. We use three variables, with the help of which, we find the sales variable. We also find the sales rate amongst people of various age groups, gender and cities.

The sales rate is dependent on the various categories that are mentioned in the dataset. We try to solve this problem by analysing the data we have from the Black Friday Sale, allowing us to make good and efficient predictions.

2 ABOUT THE DATASET–

The dataset here is a sample of the transactions made in a retail store on a Black Friday. The store wants to know better the customer purchase behaviour against different products.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

The training data consists of 537577 observations and 12 features.
For pre-processing, we inspect each feature of the data for the following reasons:1)removal of features with regular missing data or assigning the value zero wherever appropriate 2)removal of unimportant, irrelevant or duplicate features. After removal of the null values, we had 164278 observations and 12 features. We then break the data into train, validation and test sets.

2.1 HYPOTHESIS GENERATION–

The following step to solve a analytics problem is to list down a set of hypothesis. In our case, it is the factors that will affect the sales on Black Friday. We predict the purchase variable with the help of three product categories. 1)Age of customer: It is only between a range of age group that people are financially stable, thereby leading to higher sales value 2)City: The culture and back ground of people vary from city to city 3)Gender: Differences in choices when gender comes in, plays a huge role.

3 METHODS USED–

3.1 MULTIPLE LINEAR REGRESSION–

It attempts to model the relationship between two or more explanatory variables and a response variable by fitting a 'linear equation' to observed data. Every value of the independent variable x is associated with a value of the dependent variable y, i.e we must identify a straight line that best fits the data. We must keep this in mind that it is only against the best fit line that we have minimum prediction errors. The R2 score for Mutiple Regression model is 0.42..

3.2 KNN REGRESSION–

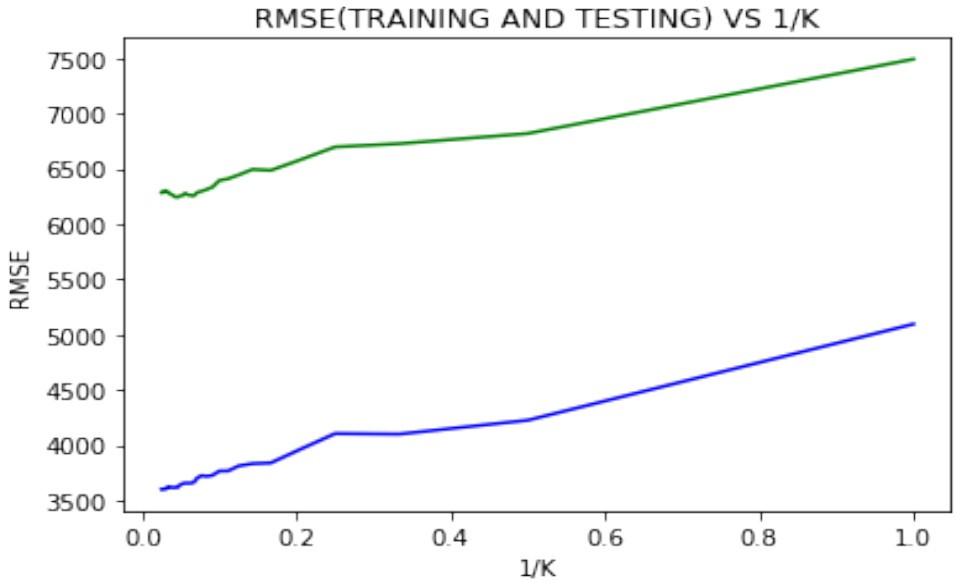
K Nearest Neighbours algorithm (KNN) is a non-parametric method used for classification and regression. In KNN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbours. The R2 score for KNN Regression model is 0.5002 and the RMSE value is 3599.63…

4 RESULTS–

Our project link is :

`https://github.com/Rahul1582/Linear-And-KNN-Regression/blob/master/Black%20Friday.ipynb`

4.1 FIGURES-



(3).png

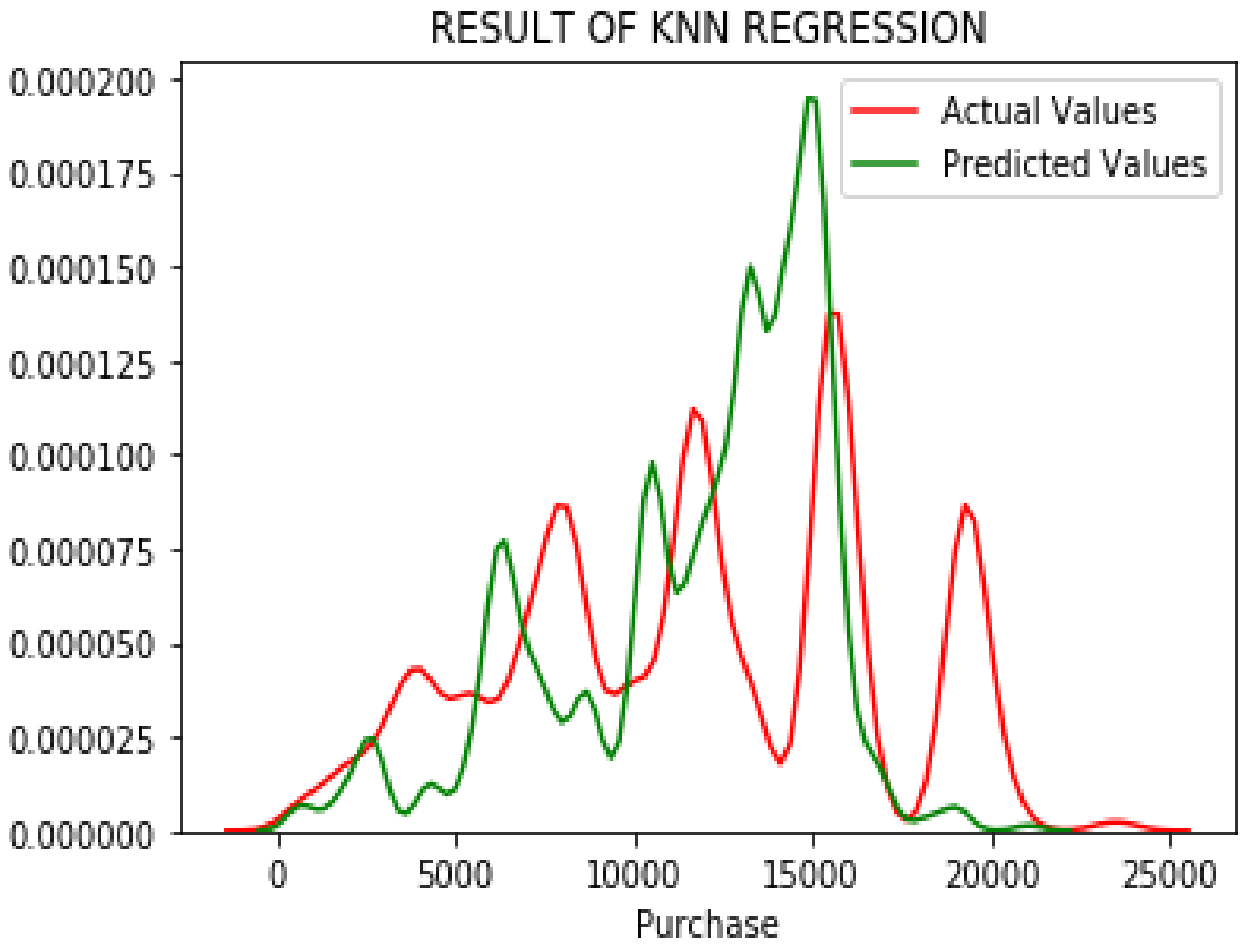
5 DISCUSSIONS–

As already stated before, we applied Multiple Linear Regression and KNN regression from which we got several values that can be used in the analysis and prediction, them being R2 score, RMSE and accuracy percentage. These values are different for Multiple Linear Regression and KNN regression and are given in the table below:

Table 1: Linear Regression VS KNN Regression

(R2 SCORE , RMSE AND ACCURACY)			
	R2 Score	RMSE	Accuracy(R2*100)
Linear Regression	0.42	4639.26	42 Percent
KNN Regression	0.5002	3599.63	50.02 Percent

(7).png



6 CONCLUSIONS–

Several Machine Learning techniques like Multiple Linear Regression and KNN Regression was used. After having applied both Multiple Linear Regression and KNN Regression, we observe that KNN Regression worked better than Multiple Linear Regression in terms of prediction efficiency and accuracy.

Significant analysis was made with the help of Root Mean Squared Error, Mean Absolute Error and R2 score. In this project, we tried to come up with the best model for the prediction of sales.

Another thing that we observed is that with city, gender and age, the sales rate or the number of purchases made varies. Highest purchases were made in City B, by Males and by the people of the age group 26-35 respectively. The accuracy we acquired is creditable considering the fact that the dataset was large with several errors, hidden and null values. We must also consider that all the variables which are there in the dataset are not strongly co-related with the purchase variable.

7 ACKNOWLEDGEMENT–

We would like to thank our mentor Ankit Tewari, who guided us on our first project, of as to how to apply Multiple Linear Regression and KNN regression, even while handling datasets with a large number of fields and while solving confusing problems which contains several qualitative variables.

References

[1] Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani: An Introduction to Statistical Learning with Apllications in R ..

[2] https://www.kaggle.com/mehdidag/black-friday

[3] https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[4] http://www.statsoft.com/Textbook/Multiple-Regression