

Black Friday Sales Prediction Project

Use Case Report



Submitted by:

Kirti Ohlan

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Black Friday Sales Prediction, which also helped me in doing lots of research wherein I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1. <https://www.google.com/>
2. <https://www.youtube.com/>
3. <https://scikit-learn.org/stable/user_guide.html>
4. <https://github.com/>
5. <https://www.kaggle.com/>
6. <https://medium.com/>
7. <https://towardsdatascience.com/>
8. <https://www.analyticsvidhya.com/>

**INTRODUCTION**

* Business Problem Framing

The shopping sector has greatly evolved due to the Internet revolution. Most of the population takes into consideration online shopping more than the traditional method of shopping. The biggest perks of online shopping are convenience, better prices, more variety, easy price comparisons, no crowds, etc. The pandemic has boosted online shopping. Though online shopping keeps growing every year, the total sales for the year 2021 are expected to be much higher [16]. Black Friday originated in the USA and is also referred to as Thanksgiving Day. This sale is celebrated on the fourth Thursday of November once every year. This day is marked as the busiest day in terms of shopping. The purpose of organizing this sale is to promote customers to buy more products online to boost the online shopping sector. The prediction model built will help to analyze the relationship among various attributes. Black Friday Sales

Dataset is used for training and prediction. Black Friday Sales Dataset is the online biggest dataset and the dataset is also accepted by various e-commerce websites . The prediction model built will provide a prediction based on the age of the customer, city category, occupation, etc. The prediction model is implemented based on models like linear regression, ridge regression, lasso regression, Decision Tree Regressor, Random Forest Regressor. The paper further walks through various sections. Section I gives an introduction to the problem, section II illustrates the prior research done in this field, section III provides the data set description, section IV presents the proposed model, with the conclusion in the last section.

* Review of Literature

Ample research is carried out on the analysis and prediction of sales using various techniques. There are many methods proposed to do so by various researchers. In this section, we will summarize a few of the machine learning approaches. It has proposed a prediction model to analyze the customer's past spending and predict the future spending of the customer. The dataset referred is Black Friday Sales Dataset from analyticsvidhya. They have machine learning models such as Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging, and XGBoost. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. Simple problems like regression can be solved by the use of simple models like linear regression instead of complex neural network models. Odegua, Rising have proposed a sales forecasting model. The machine learning models used for implementation are K-Nearest Neighbor, Random Forest, and Gradient Boosting. The dataset used for the experimentation is provided by Data Science Nigeria, as a part of competitions

* Motivation for the Problem Undertaken

Predicting customer's behaviour is one of the most popular applications of Machine Learning in various fields like Finance, Sales, Marketing. Building such predictive models, we can predict the impact of the decisions taken on the growth of our organization.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the sales of variables & also how do these variables describe the sales. This will help to determine the sales with the available independent variables.

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the

conditional expectation of the dependent variable when the independent variables take on a given set of values.

Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

* Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 550068 rows and 12 columns. There are 2 data sets that are given. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. Size of training set: 550068 records.

2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 233599 records.

* Data Preprocessing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre- process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

1. Creating the dataset as a DataFrame.
2. Used pandas to set display I ensuring we do not see any truncated information.
3. Checked the number of rows and columns present in our training dataset.
4. Checked for missing data and the number of rows with null values
5. Duplicate values are also present in our dataset.
6. Checked the unique values information in each column to get a list for categorical data
7. Used Pandas Profiling during the visualization phase along with violin plot, count plot, cat plot and the others
8. With the help of ordinal encoding technique converted all object datatype columns to numeric datatype
9. Thoroughly checked for outliers and skewness information
10. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
11. Separate feature and label data to ensure feature scaling is performed avoiding any kind of biasness
12. Checked for the best random state to be used for our Regression Machine Learning model pertaining to the feature importance details
13. Finally created a regression model function along with evaluation metrics to pass through various model formats

* Data Inputs- Logic- Output Relationships

When we loaded the dataset, we had to go through various data pre-processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to logical part the domain expertise of understanding how real estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In Data Science community there is a saying “Garbage In Garbage Out” therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data how they were related to each other as well as our target label.

* With the objective of predicting flight prices accurately we had to make sure that a model was built that understood the customer priorities trending in the market imposing those norms when a relevant price tag was generated. I tried my best to retain as much data possible that was collected. I did not want to impute data and then cause a biasness in the machine learning model from values that did not come from real people.
* Hardware and Software Requirements and Tools Used

Hardware Used:

1. RAM: 8 GB
2. CPU: AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz
3. GPU: AMD Radeon ™ Vega 8 Graphics and NVIDIA GeForce GTX 1650 Ti

Software Used:

1. Programming language: Python
2. Distribution: Anaconda Navigator
3. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn and pandas profiling

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models.

For this project we need to predict the sales, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected K Neighbors as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of less amount of data.

In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.

Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted price values out of the Regression Machine Learning Model.

* Testing of Identified Approaches (Algorithms)

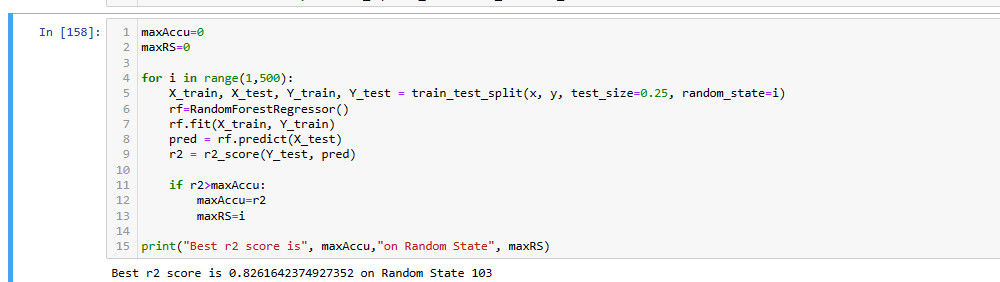
The algorithms used on training and test data are as follows:

1. Linear Regression Model
2. Ridge Regularization Regression Model
3. Decision Tree Regression Model
4. Random Forest Regression Model
5. K Nearest Neighbours Regression Model

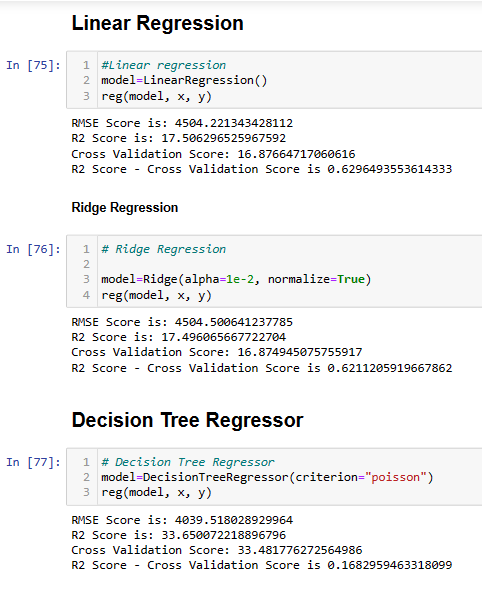
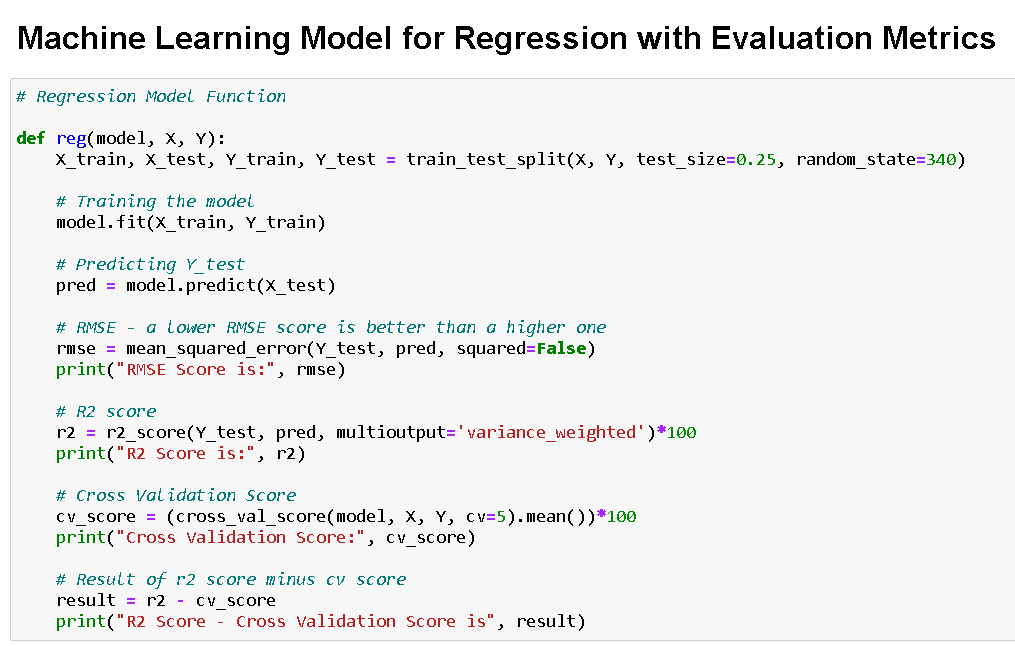
* Run and Evaluate selected models

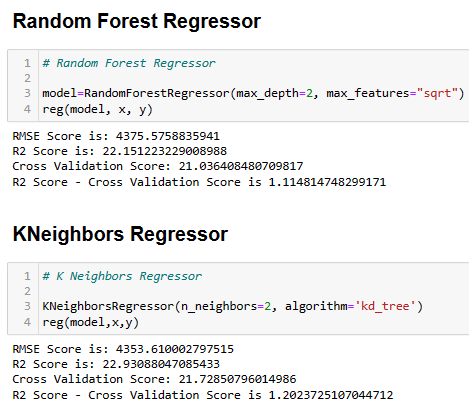
I used a total of 5 Regression Models after choosing the random state amongst 1-500 number. Then I even defined a function for getting the regression model trained and evaluated. The code for the models is listed below.

Random State



Regression Model Function:





* Key Metrics for success in solving problem under consideration

The key metrics used here were r2\_score, cross\_val\_score, MAE, MSE and RMSE. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

1. Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces

where each part being 20% of full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

2. R2 Score:

It is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

3. Mean Squared Error (MSE):

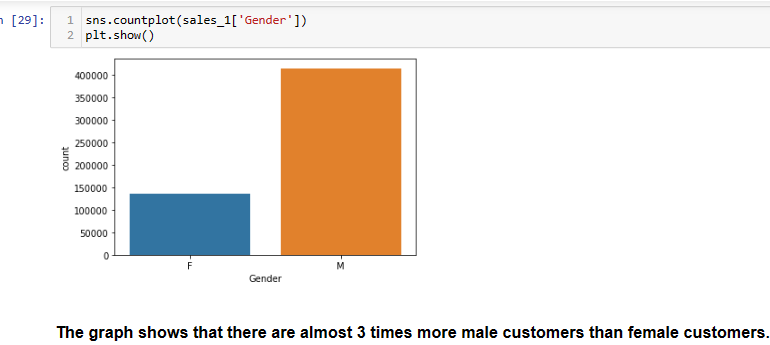
MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. RMSE is the Root Mean Squared Error.

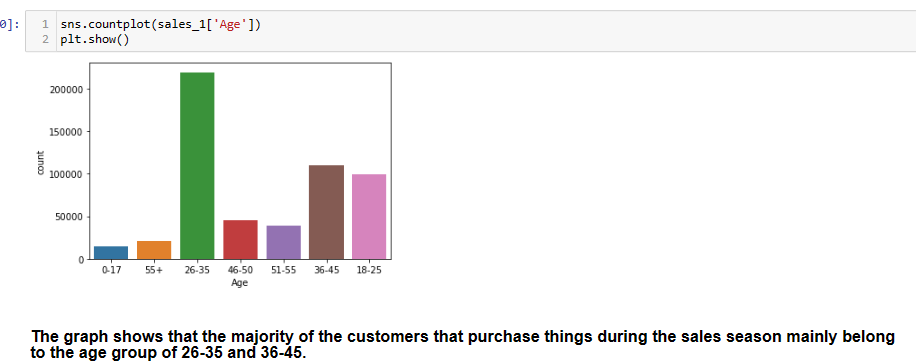
4. Mean Absolute Error (MAE):

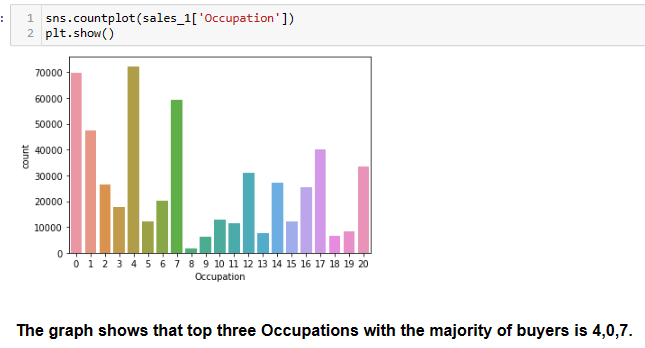
MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

* Visualizations

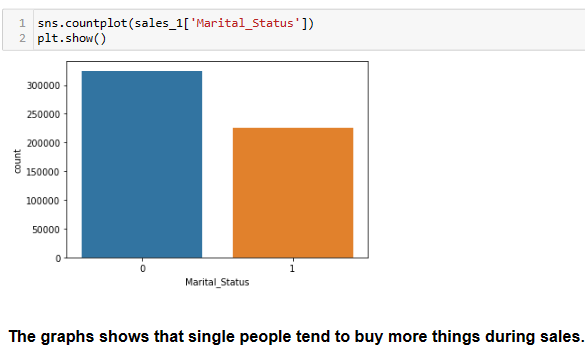
I created cat plots, count plots and violin plots to get further visual insights on our dataset feature value.

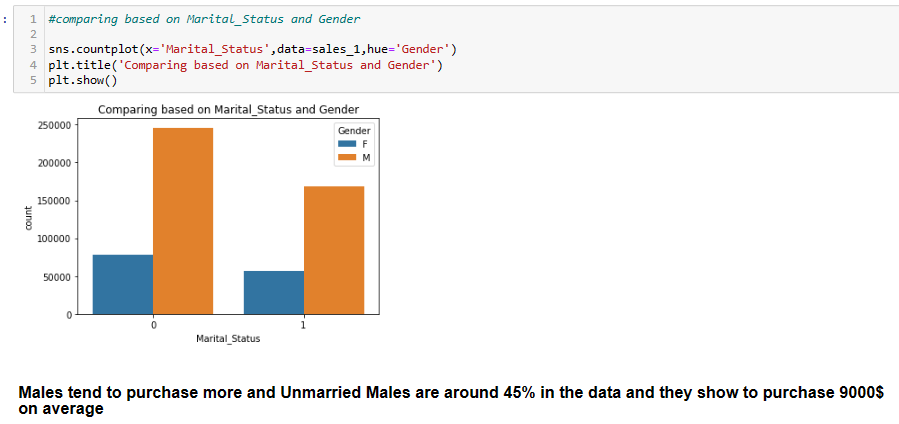


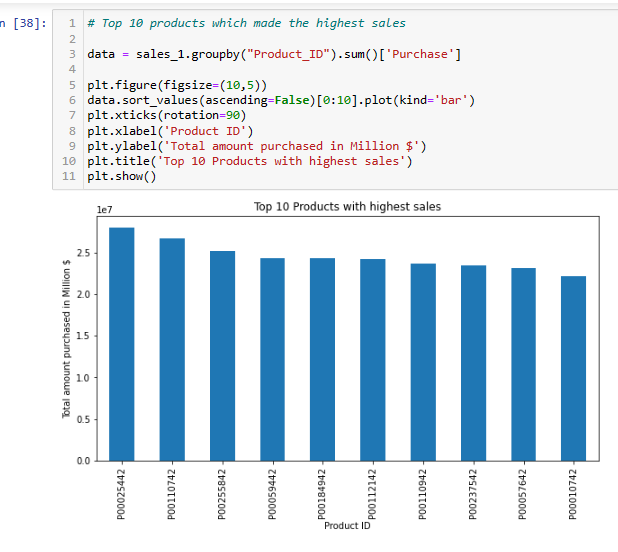


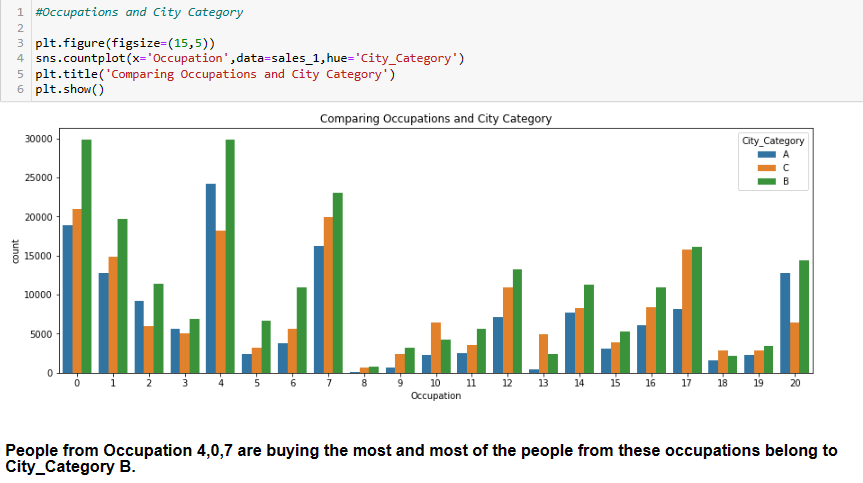












* Interpretation of the Results

Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importance and to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot. I got to know the count of a particular category for each feature by using count plot and most importantly with predicted target value distribution

Pre-processing: Basically, before building the model the dataset should be cleaned and scaled by performing few steps. As I mentioned above in the pre-processing steps where all the important features are present in the dataset and ready for model building.

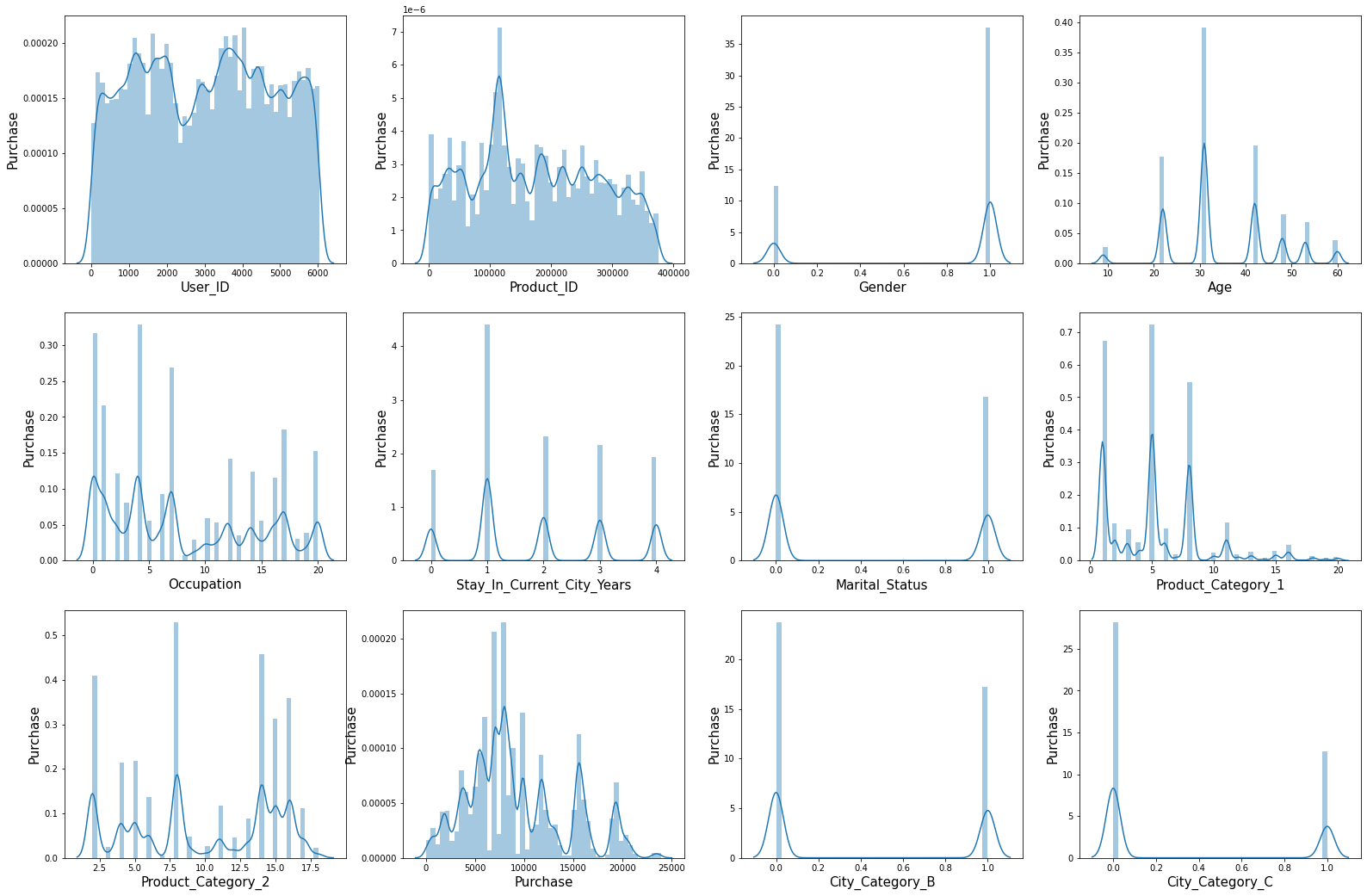
Model Creation: Now, after performing the train test split, I have x\_train, x\_test, y\_train&y\_test, which are required to build Machine learning models. I have built multiple regression models to get the best R2 score, MSE, RMSE & MAE out of all the models.

**CONCLUSION**

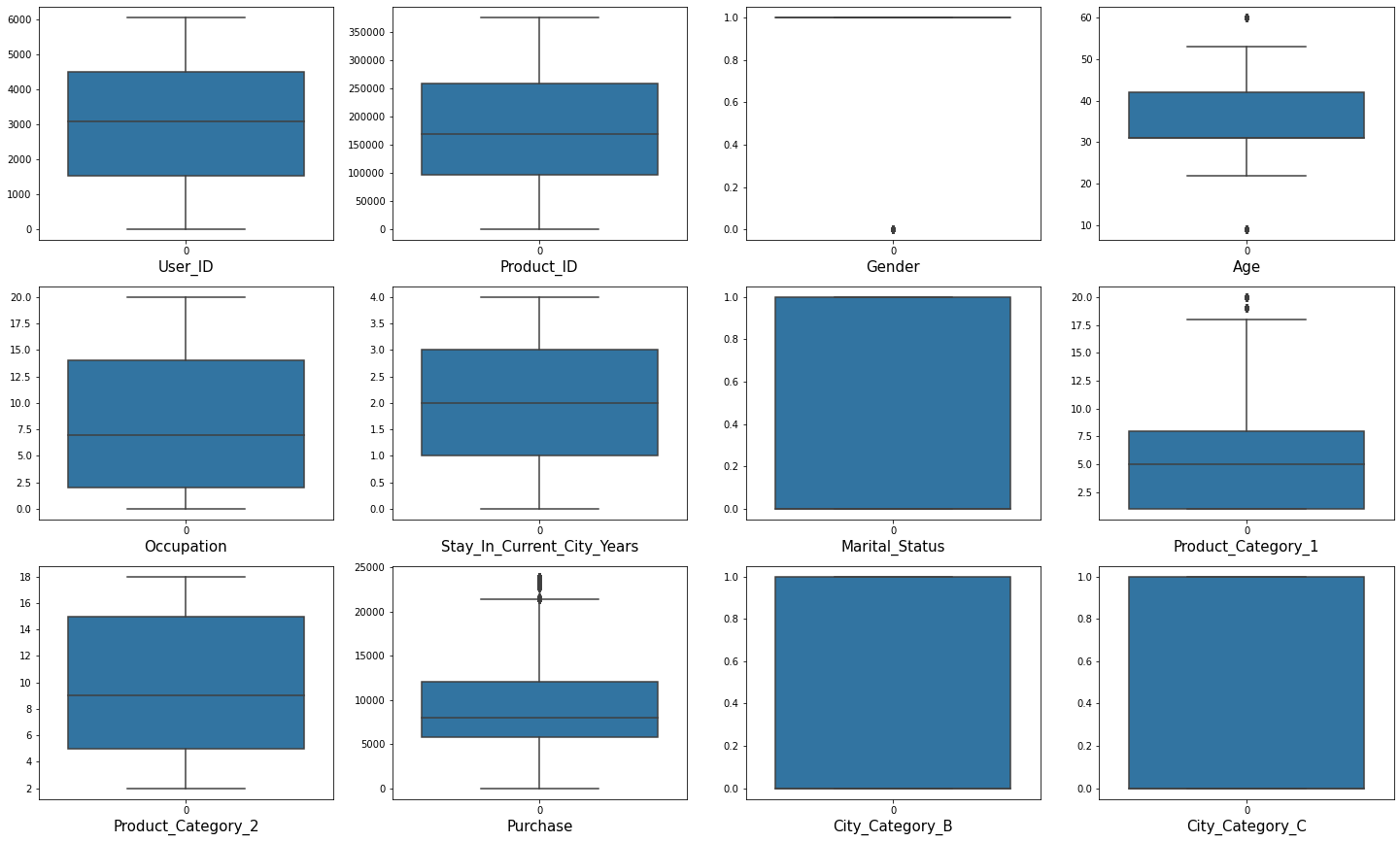
* Key Findings and Conclusions of the Study

I observed all the encoded dataset information by plotting various graphs and visualised further insights.

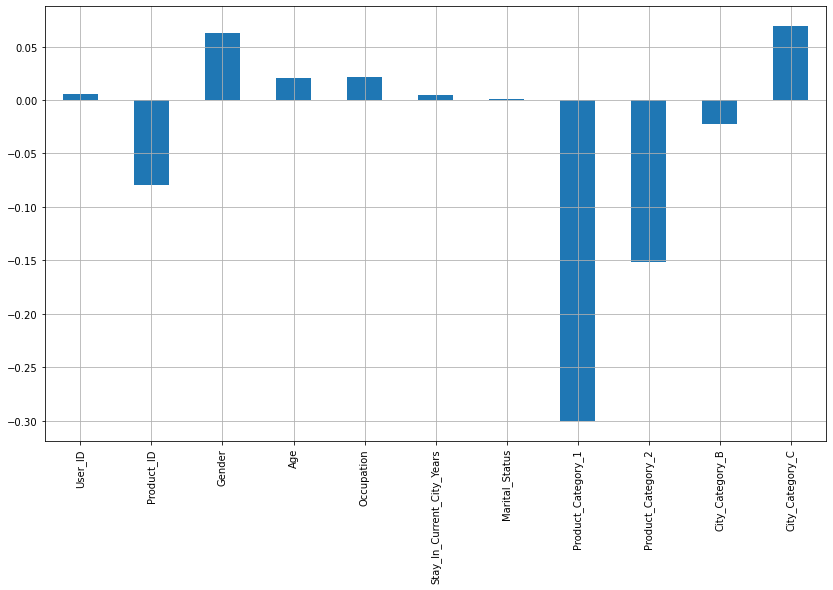
Distplot:



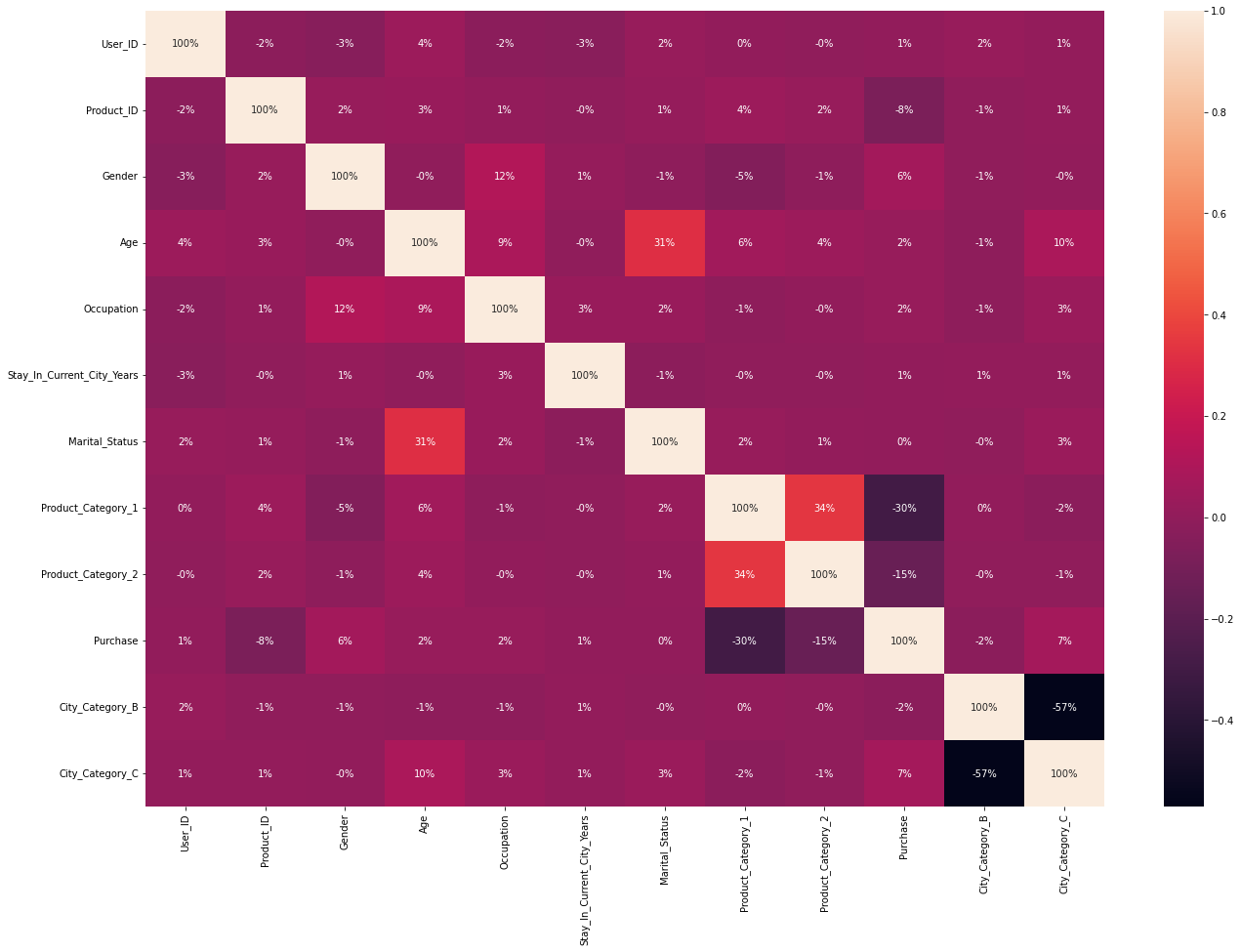
Boxplot:



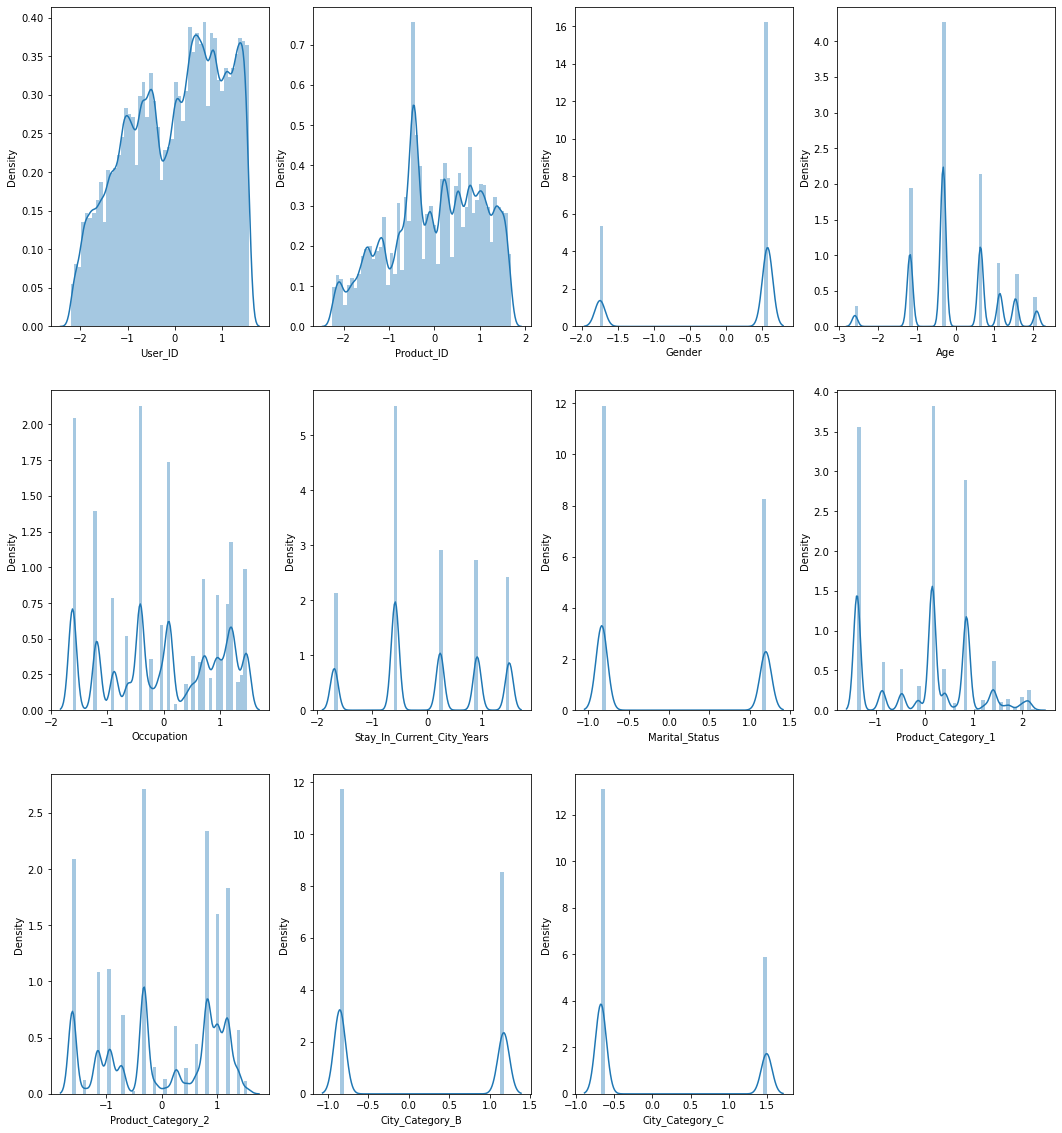
Correlation:



Heatmap:



Distribution plot:



**Conclusion**

With traditional methods not being of much help to business growth in terms of revenue, the use of Machine learning approaches proves to be an important point for the shaping of the business plan taking into consideration the shopping pattern of consumers. Projection of sales concerning several factors including the sale of last year helps businesses take on suitable strategies for increasing the sales of goods that are in demand. Thus the dataset is used for the experimentation, Black Friday Sales Dataset . The models used are Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regressor, and Random Forest Regressor. The evaluation measure used is Mean Squared Error (MSE). Based on Table II Random Forest Regressor is best suitable for the prediction of sales based on a given dataset. Thus the proposed model will predict the customer purchase on Black Friday and give the retailer insight into customer choice of products. This will result in a discount based on customer-centric choices thus increasing the profit to the retailer as well as the customer.