PROJECT PAPER

Black Friday Purchase Prediction

Devesh Sharma (664049903) Nimisha Asati (678028458) Shadabi (659548611)

ABSTRACT:

Performed various modelling techniques like Linear Regression, Decision Tree, XG Boost and SVM to predict sales during Black Friday of a retail store. The best accuracy was given by Naïve Bayes followed by Decision Tree. We also performed K-means clustering to achieve clusters consumable by the business. We were successful in getting 3 distinct clusters which showcased similar characteristics among themselves and heterogenous behavior with each other. This analysis helps the store predict how much sales can they expect during Black Friday season and how to segment their customers more effectively in order to target them from promotional campaigns.

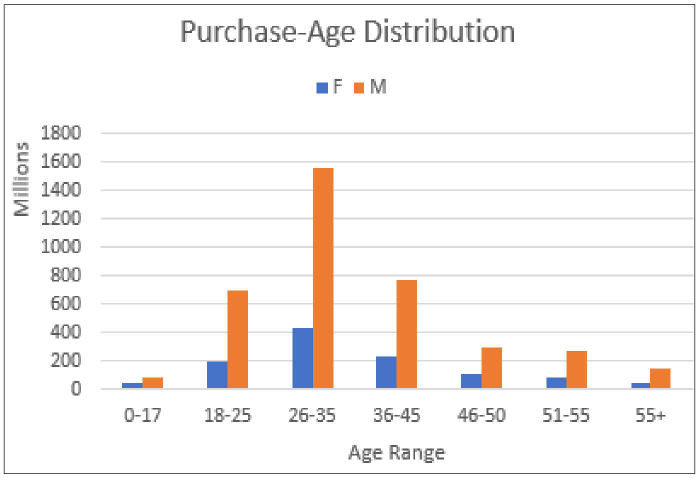
INTRODUCTION:

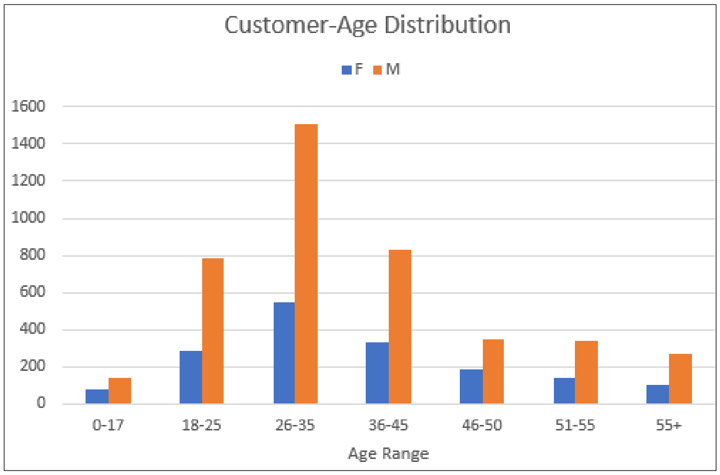
The problem in hand is a study of sales through customer behavior in a retail store during Black Friday. The store wants to understand the customer purchase behavior better against different products. This problem was interesting because it offered us scope to try out multiple statistical techniques, solving a variety of problem like regression, classification and clustering.

Practical usage could be deployment of the model in any retail store to predict sales during Black Friday season and for segmentation purpose, to better design promotional campaigns.

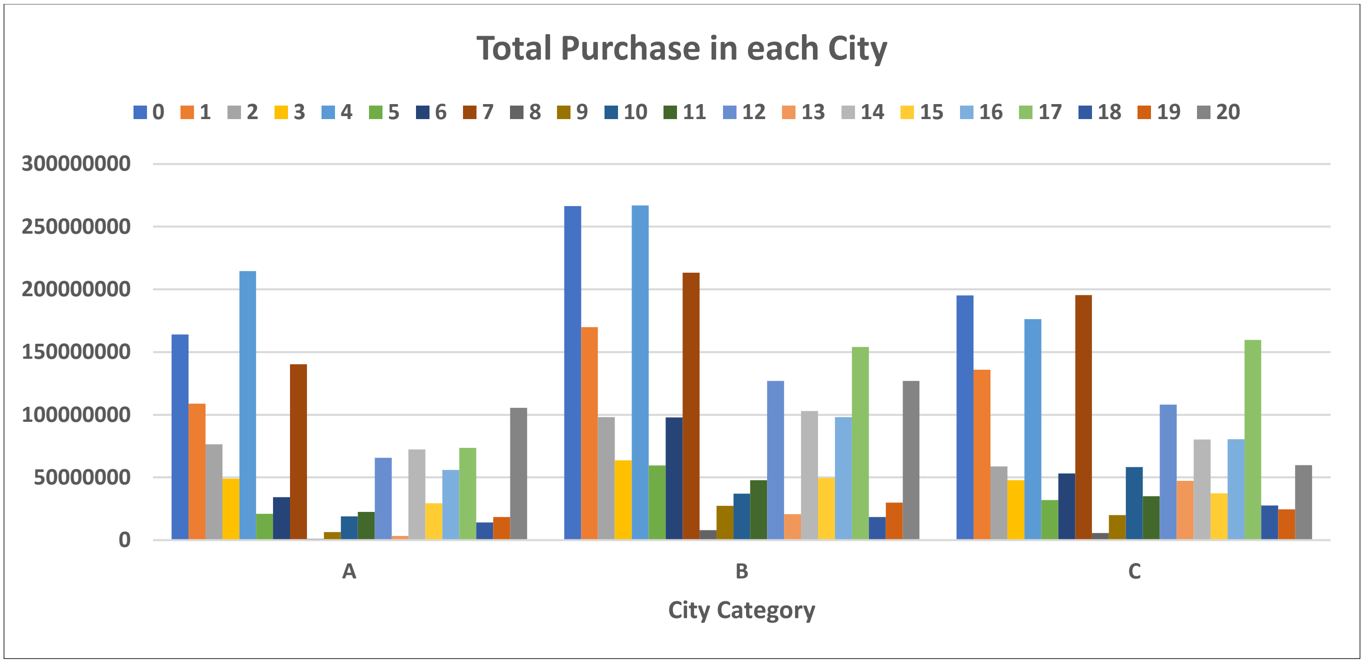
The dataset here is a sample of the transactions made in a retail store (picked from Kaggle). We have around 530K examples and 12 features which are all categorical. Upon initial glancing through the data and keeping the business requirement in mind, we can assume that variables like Gender, Age, Occupation and Product Category would be important towards prediction of sales.

Exploratory Data Analysis:





The maximum purchases were made by age group 26-35. And the male spent predominantly more than female.



The above plot shows distribution of purchase patterns for different cities based on the occupations. We can infer that; the highest purchasing population resides in city B whose occupations are 0 and 4.

RELATED WORK:

We referred to the research paper on Diagnosis of Breast Cancer using Decision Tree Models and SVM, where they have both categorical and continuous variables. They performed decision tree and radial basis function kernel support vector machine to assign patients to benign or malignant breast cancer group. They choose the Gini index, which is based on probabilities of category membership in the branch and built a tree, to create subgroups with similar output values and minimize the impurity within each node. The SVM model gave the best performance for their model. The classification accuracy improved as they increased the width of the radial basis function, but this can also lead to overfitting. They performed comparative analysis for different decision tree models which were Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression tree (C&R), Quick Unbiased Efficient Statistical Tree (QUEST), and Ross Quinlan new decision tree model C5. 0 and RBF-SVM.
Similarly, we pre-processed our data and performed transformations, and trained the model using decision tree, RBF-SVM and other methods. Our model showed similar results and did not perform well on linear regression. The RBF-SVM showed the best performance with highest accuracy for our model.

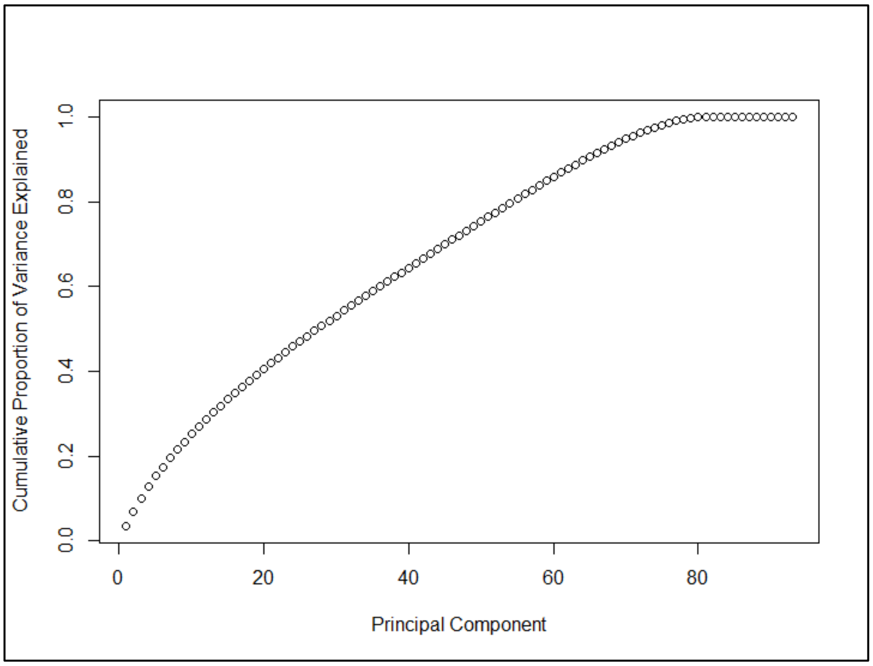
MODELS AND METHODS:

Handling Missing Values: We neither could remove the missing values, given their high frequency in the dataset nor replace the missing values with their mean value, given the categorical nature of the variables. Therefore, we replaced all the missing values with the mode value. This imputation can introduce bias in our data but will provide us the ‘most likely’ prediction.

Data Transformation: For some of the models, we would be required to have numerical variables. Therefore, we transformed our categorical variables into dummy numerical variables, each with 1/0 flag based on the occurrence of that value for the concerned row. Upon this transformation, we achieved 96 binary variables, as compared to 12 variables at the start.

Correlation with the Target Variable: To analyze the impact of independent variables on our target ‘purchase’ variable, we used correlation matrix by converting categorical data into their numerical counterparts. We then deduced the variables having high correlation with the target variable (both positive and negative). The result resonated with our initial finding, product category, age and gender being one of the highest correlated variables.

Principal Component Analysis: We performed PCA to reduce the number of regressors such that the resulting ones capture the maximum variance of the entire dataset. On observing the variance graph, we could deduce that between 70-80 variables were required to capture enough cumulative variance of the entire dataset.



Models and Parameter Estimation:

1. Linear Regression: Utilizing the highly correlated variables to predict the relation between purchase amount and various other regressors. The underlying assumption in this scenario is that these regressors would have linear relationship with the target variable.
   Parameters: We used above obtained principal components in the model, making sure that the inter variable VIF value is low, ensuring low multicollinearity. Training and Test data was created by splitting the original dataset into 70:30, 80:20 or 60:40 ratio.

2. Decision Tree: We built decision tree models considering all the categorical variables, without transforming them into binary dummy variables. Also, transformed the target variable into factor variable, by placing its values in bins achieving 5 different classes.
   Parameters: We used Gini and Information Gain as splitting parameter, capping maximum depth of the tree and using class method. Training and Test data was created by splitting the original dataset into 70:30, 80:20 or 60:40 ratio.

3. Naïve Bayes: We developed a basic Naïve Bayes model in order to understand what each feature is contributing independently towards the probability of the target variable.
   Parameters: Training and Test data was created by splitting the original dataset into 70:30, 80:20 or 60:40 ratio.

4. XG Boost: We ran XG boost models with cross validation on the dataset by replacing all missing values with mode and transforming the variables into their binary counterpart. Target variable is also transformed to have buckets (ranges of values).
   Parameters: We tried combinations of nrounds, nfold and num class while keeping other parameters like booster tree and evaluation metric constant. Training and Test data was created by splitting the original dataset into 70:30, 80:20 or 60:40 ratio.

5. SVM: We implemented SVM on the training dataset using both Linear and Radial kernels. For this also, we transformed the categorical variables into their numerical dummy versions and then fed into the model.
   Parameters: We used gamma values ranging from 0.01 to 0.05, in intervals of 0.01. Changed kernel type to both linear and radial. Training and Test data was created by splitting the original dataset into 70:30, 80:20 or 60:40 ratio.

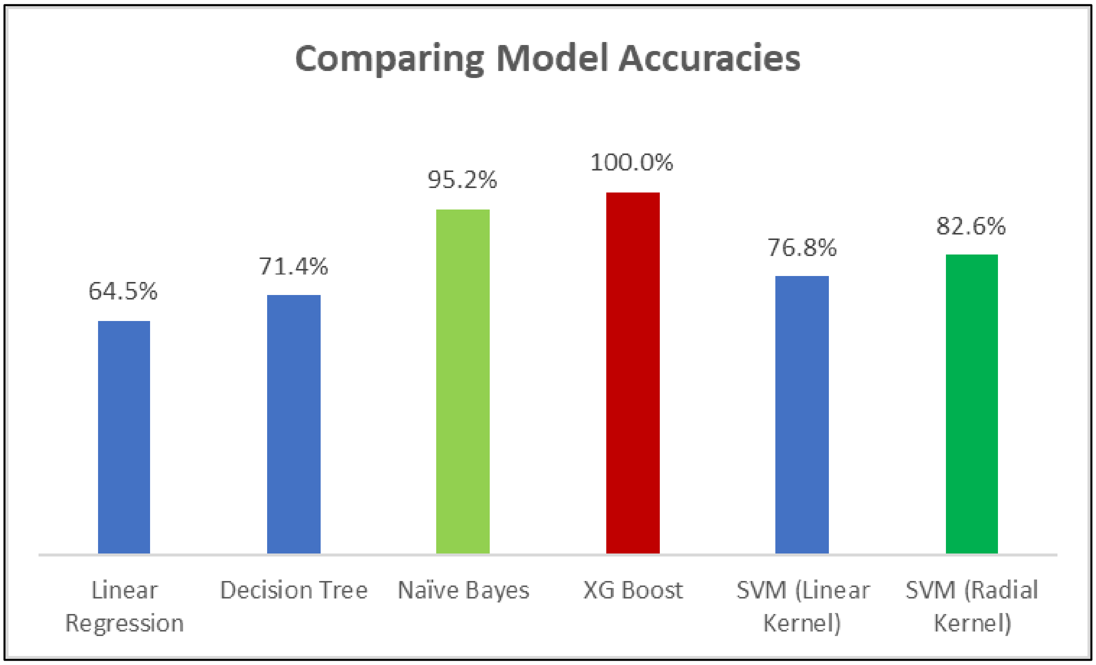
6. K-means clustering: In the end, we also ran K-means clustering algorithm on our dataset to gain the commercially viable clusters, having business implications for the client. Here also, we transformed the categorical variables into their binary dummy variables for the algorithm to calculate distance between data points.
   Parameters: We used K value as 3, 4 and 5 based on the Elbow curve analysis and also selected maximum iterations as not more than 10.

EXPERIMENTAL RESULTS:

Model Performance Comparison: Here, we are comparing the best model obtained under each category (in terms of overall accuracy).

Model	Parameters	Accuracy
Linear Regression	Split Ratio -> 70:30	64.5%
Decision Tree	Split Ratio -> 70:30	71.4%
	Split Parameter -> Gini	
	Max Depth -> 10	
	Method -> Class	
Naïve Bayes	Split Ratio -> 70:30	95.2%
XG Boost	Nrounds -> 50	100%
	Nfold -> 10	
	Num class -> 5	
	Booster -> gbtree	
	Split Ratio -> 70:30	
SVM	Kernel -> linear	76.8%
	Gamma (as list) -> 0.01, 0.02, 0.03, 0.04 and 0.05	
	Split Ratio -> 70:30	
SVM	Kernel -> radial	82.6%
	Gamma (as list) -> 0.01, 0.02, 0.03, 0.04 and 0.05	
	Split Ratio -> 70:30	

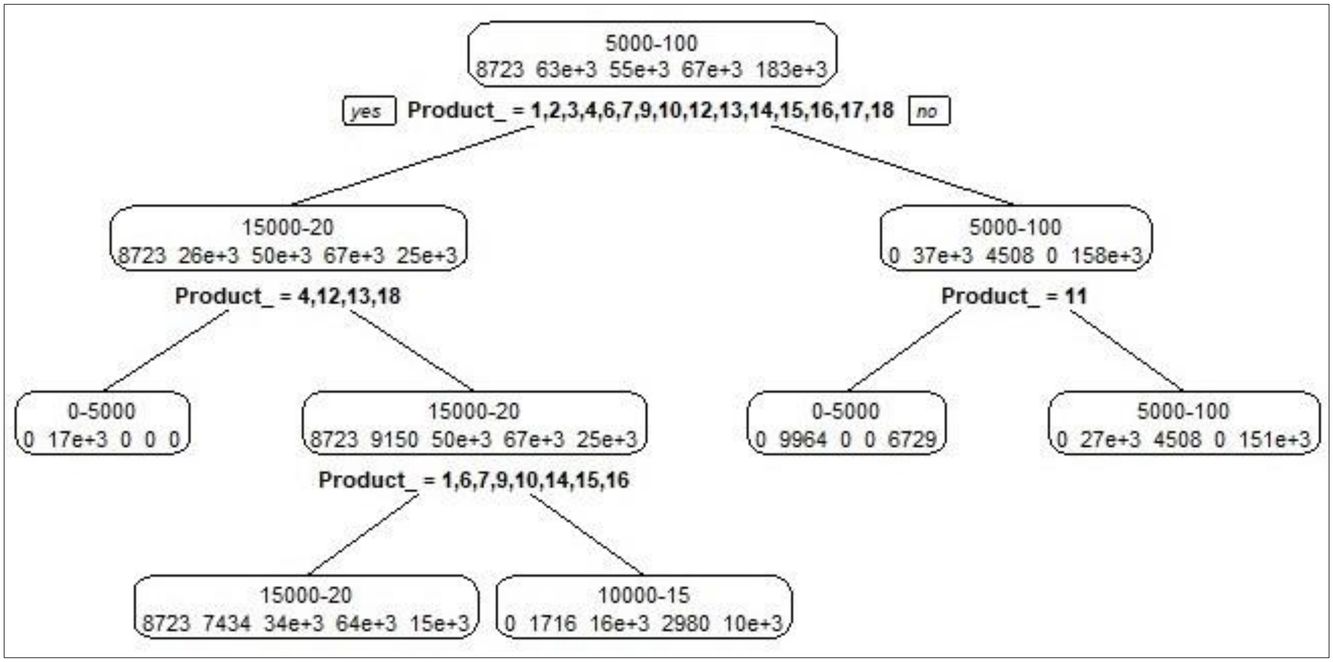
We can observe that Linear regression performed the worst among all whereas Naïve Bayes and radial basis SVM performed the best among all. XG Boost gave 100% accuracy but it is not practical to predict all instances correct, therefore we cannot select this model. Naïve Bayes acts as the baseline model, also giving here the highest accuracy (which is also a bit on the higher side).



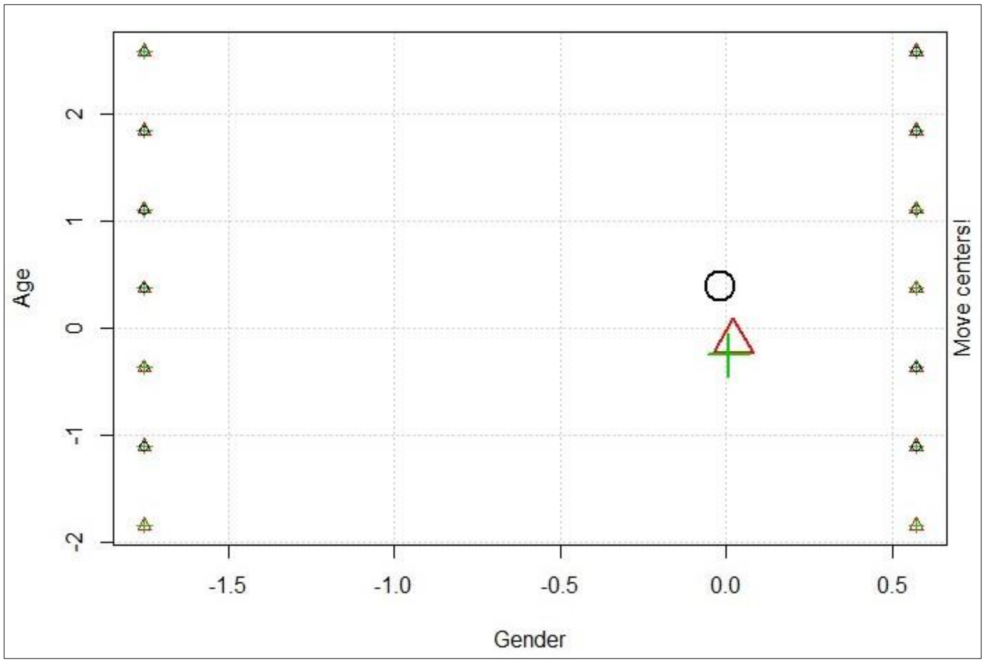
Have highlighted XG Boost as red because it was not giving practical result. Radial Based SVM model gave the best result while showcasing a good generalization, followed by Naïve Bayes.

Model Plots:

Decision Tree



K-means Clustering



DISCUSSION:

When the data under consideration is entirely categorical, it gets difficult for the linear regression algorithm to predict the target variable. It is only when we have few continuous variables to regress, that we can expect linear regression technique to work and therefore our linear regression model was not giving very good result.

When we compare Decision Tree and Naïve Bayes models as classification technique, Decision Tree might lead to overfitting of the model whereas that is not the case with Naïve Bayes. Also, Decision Tree works best when we do not have multiple classes in our target variable. This got proved by our modelling result also when Naïve Bayes gave a better accuracy compared to Decision Tree, when our target variable has 5 different classes. Whenever there are a lot of examples but very less features to deal with, Naïve Bayes tends to perform better without overfitting.

XG Boost is scalable alternative of gradient tree boosting. This technique is very fast in dealing with ever growing big data, maintain a very high accuracy.

SVM model can be run using 2 primary kernels: Linear and Radial. Whenever, our dataset has linear relationship between target and independent variables, we can expect SVM with Linear kernel to give better results otherwise SVM with Radial kernel performs better. This is the case with our dataset also, which does not have a linear relationship between target and independent variables and therefore, Radial kernel based SVM performed better than Linear kernel.

CONCLUSION:

Based on the above-mentioned findings and analysis, we can conclude that predicting sales during Black Friday season is one of the best analytical exercise. We could implement multiple modelling techniques, still leaving a scope for further improvements and generalization with other modelling techniques. Radial based SVM model gave the best result with good generalization whereas Naïve Bayes seemed to overfit a bit, given the base model. Naïve Bayes was able to perform exceptionally because of the multiple classes available in our dataset with numerous examples but less features. Performing K-means clustering also generated viable clusters having heterogenous inter-cluster nature and homogenous intra-cluster nature. This would enable a store to segment their customers more effectively and target them better/ relevant during promotional campaigns.

Thus, this analysis helps a store to forecast the sales which it can expect during Black Friday holiday season, given the various factors impacting the same and better design their promotional strategies.

REFERENCES:

1. https://www.researchgate.net/publication/269166287_Diagnosis_of_Breast_Cancer_using_Decision_Tree_Models_and_SVM
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/
3. https://www.researchgate.net/publication/279913074_SVM_Classification_with_Linear_and_RBF_kernels