Sales Analysis

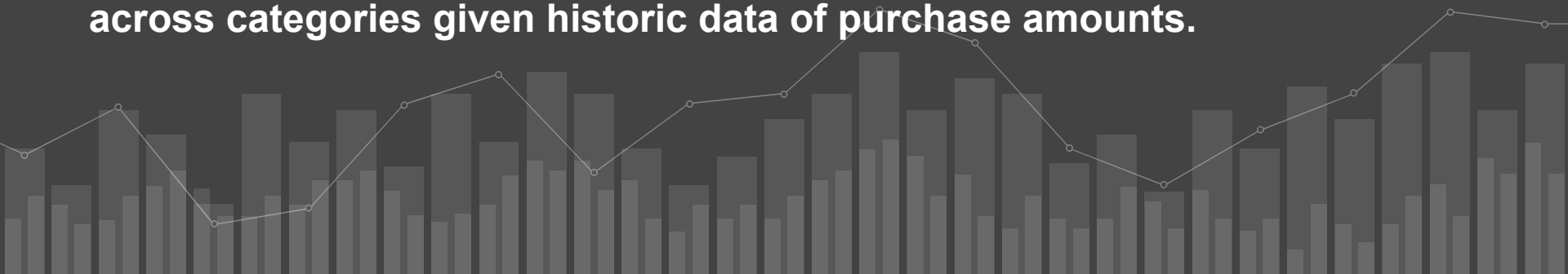
By RSA Analyzers
>Aishwarya Bhonde
>Sampada Shyam
>Rohit Raj Koul

Motivation

To help the market to focus on the factors that can boost up their sales during Black Friday.

Problem Statement

The challenge is to predict the purchase of various products by users across categories given historic data of purchase amounts.



Background and Dataset

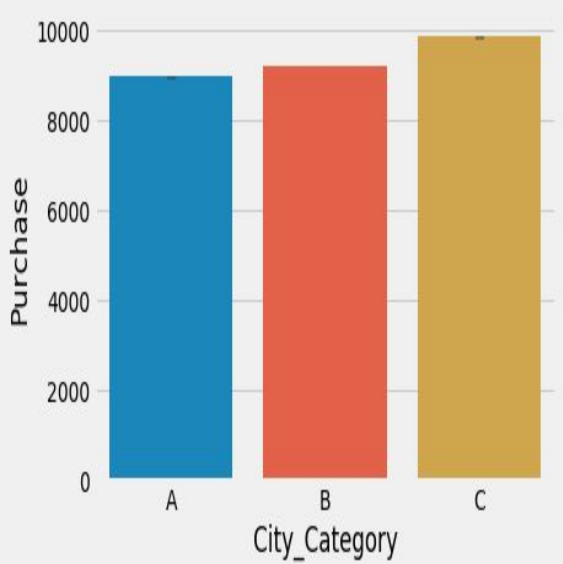
- Black Friday is the most important day for all the companies to maximize their sales and it would be very helpful to predict the sales beforehand.
- Dataset is retrieved from Kaggle.

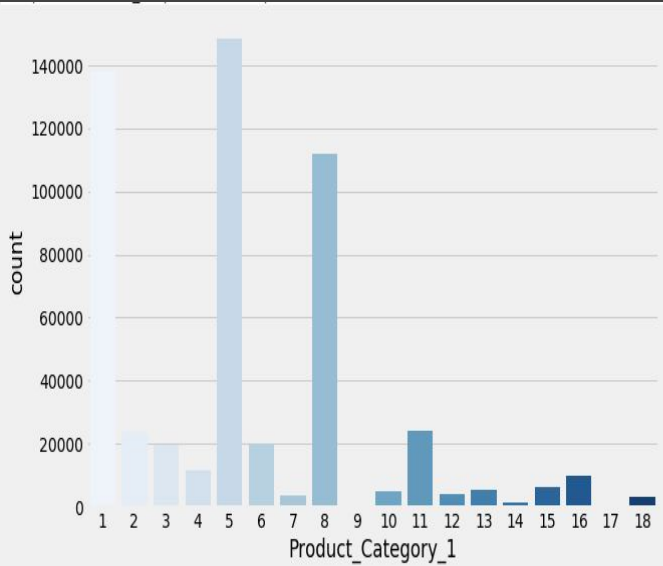
	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	NaN	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	NaN	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	NaN	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	NaN	7969
...	...	...	...	...	...	...	...	...	...	...	...	...
537572	1004737	P00193542	M	36-45	16	C	1	0	1	2.0	NaN	11664
537573	1004737	P00111142	M	36-45	16	C	1	0	1	15.0	16.0	19196
537574	1004737	P00345942	M	36-45	16	C	1	0	8	15.0	NaN	8043
537575	1004737	P00285842	M	36-45	16	C	1	0	5	NaN	NaN	7172
537576	1004737	P00118242	M	36-45	16	C	1	0	5	8.0	NaN	6875

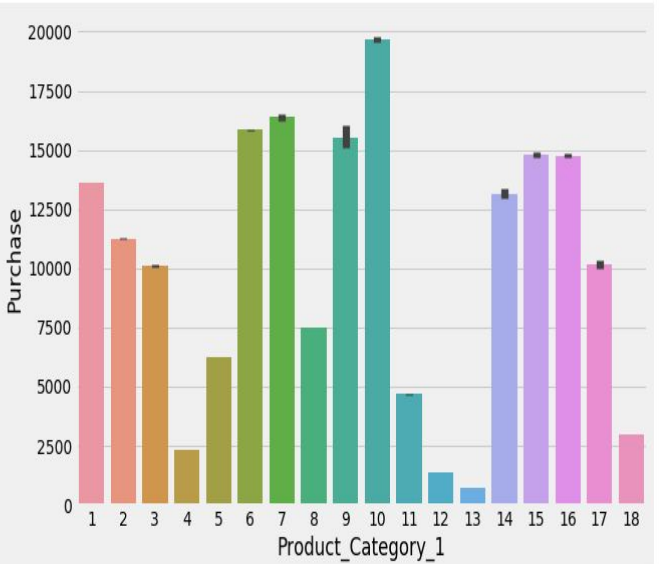
537577 rows × 12 columns

Exploratory Data Analysis

- Used at an early stage in an Analytical process.
- Explains specific sorts of initial analysis and findings done with data sets.

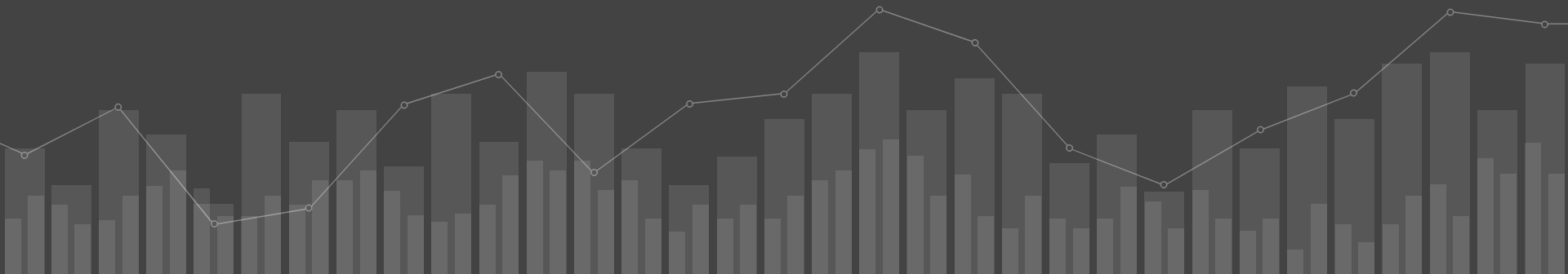






Data Preprocessing

- **Feature Selection**
- **Handling Missing Values**
- **Handling Categorical Values**
- **Splitting Data**
  Training data = 80%
  Testing data = 20%

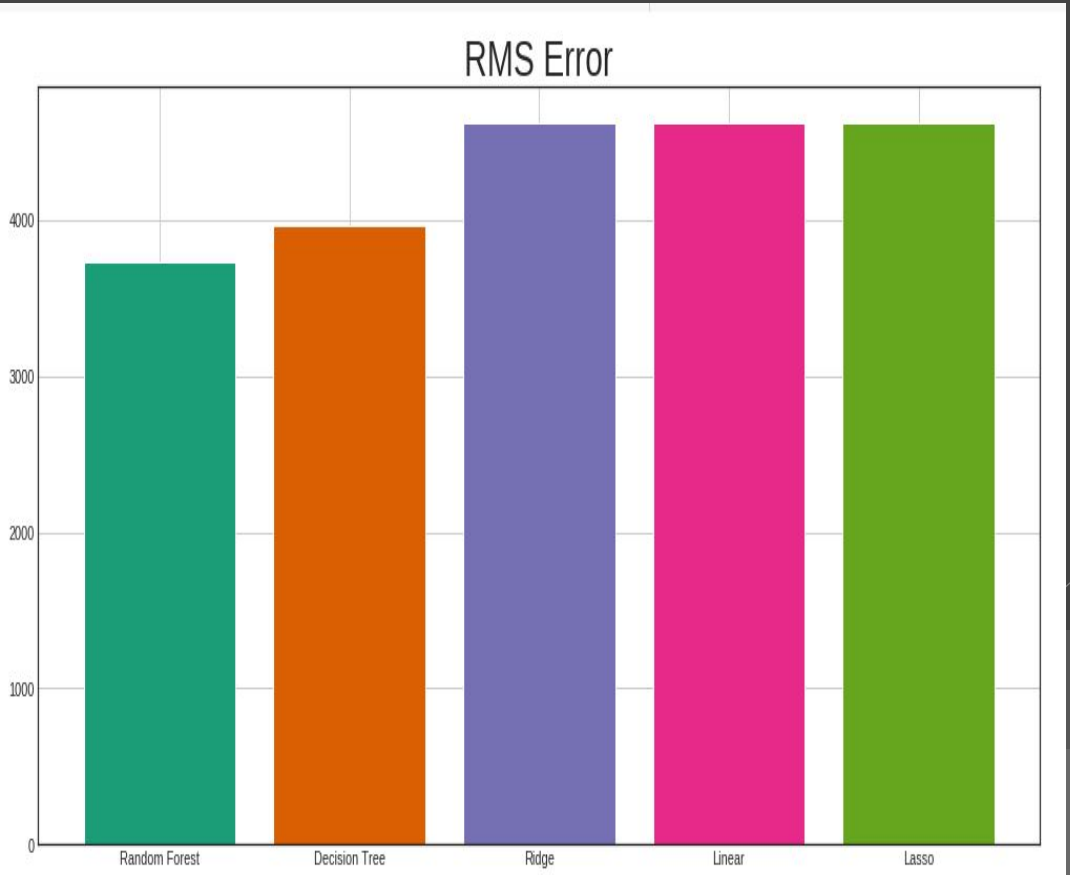


Regression Model Implementation

- **Linear Regression**
- **Ridge Regression**
- **Random Forest Regression**
- **Decision Tree Regression**
- **Lasso Regression**

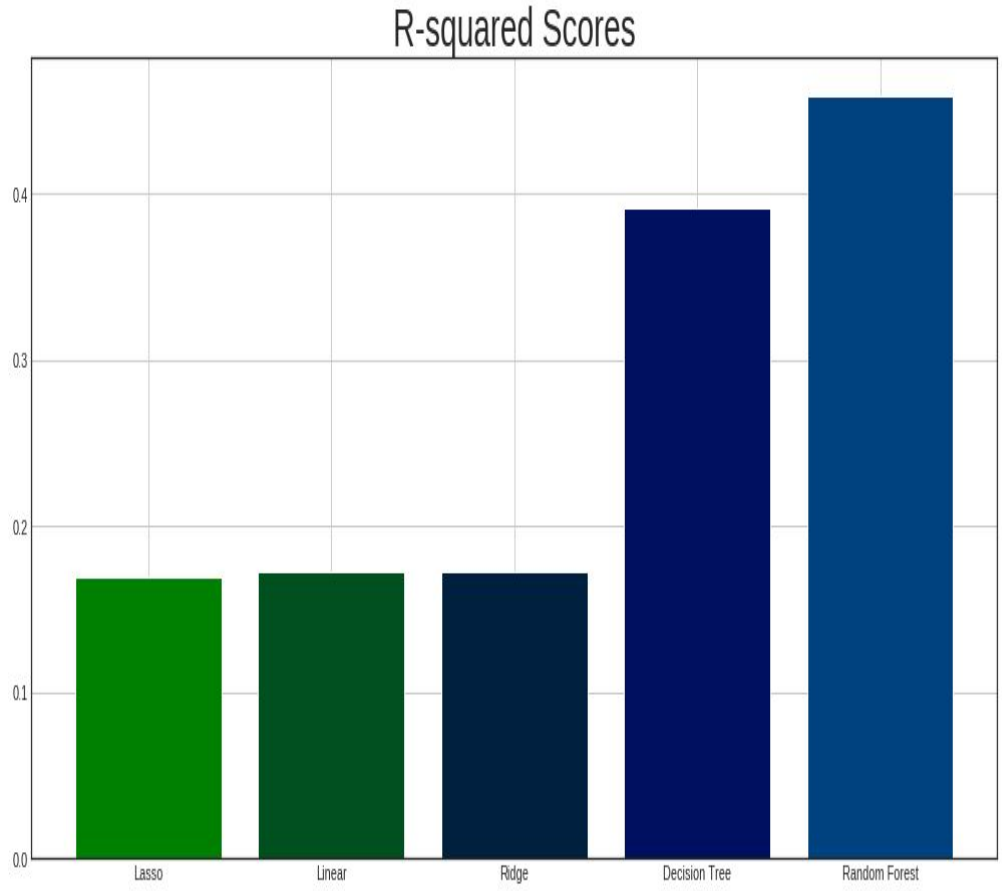
Regression Models	RMSE	R-squared scores
Linear	4619.078	0.1722
Ridge	4619.078	0.1722
Random Forest	3729.843	0.4602
Decision Tree	3960.993	0.3912
Lasso	4627.025	0.1693

Comparative Analysis of RMS Errors



- Root Mean Square Error (RMSE) is the standard deviation of the residuals (gauge botches).
- Residuals are an extent of how far from the backslide line data centers are; RMSE is an extent of how spread out these residuals are.
- Root mean square error is consistently used in climatology, deciding, and regression analysis to check preliminary outcomes.
- Comparing the root mean square error of the five algorithms, Random Forest is found to have less root mean square error.

Comparative Analysis of R-squared scores



- The coefficient of assurance, indicated R2 or r2 and articulated "R squared", is the extent of the change in the reliant variable that is unsurprising from the free variable(s).
- It gives a proportion of how all around watched results are repeated by the model, in light of the extent of all out variety of results clarified by the model.
- R-squared score for random forest is observed to be the highest i.e 46% hence providing less RMS errors.
- Decision tree is observed to be the second highest with R-squared score of 39%.
- Linear and ridge regression models are having identical R-squared score of 17% which are the least of all.

Challenges

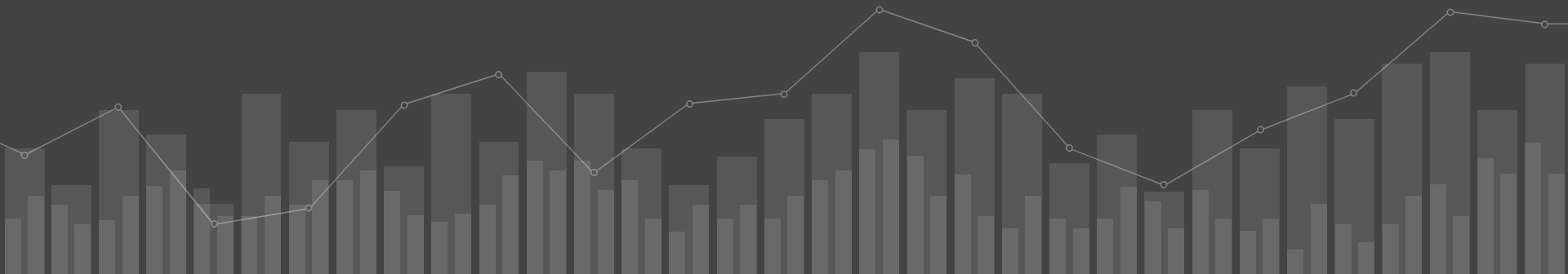
- Training Large Data Set
- Handling Missing Values
- Less Number of Features for Analysis

Future Work

- Implementing Logistic Regression, and Classification models
- Recommender System
- Adding more features like states and taxes

References

1. https://www.google.com/url?q=https://github.com/mathubhalan/Black-Friday-Sales/blob/master/Data/BlackFriday.csv



Individual Contribution:

1. Sampada Shyam
- Data pre-processing
- Linear regression model implementation

2. Aishwarya Bhonde
- Exploratory data analysis
- Ridge regression and random forest model implementation

3. Rohit Raj Koul
- Exploratory Data Analysis
- Lasso regression and decision tree classifier model implementation

Thankyou..

Questions?