Machine Learning Project

Black Friday Sales Prediction

Group 7 - DSAI K65 - HUST

July 7, 2022

Our Team Members

Nguyen Quang Duc
20204876

Le Hong Duc
20204874

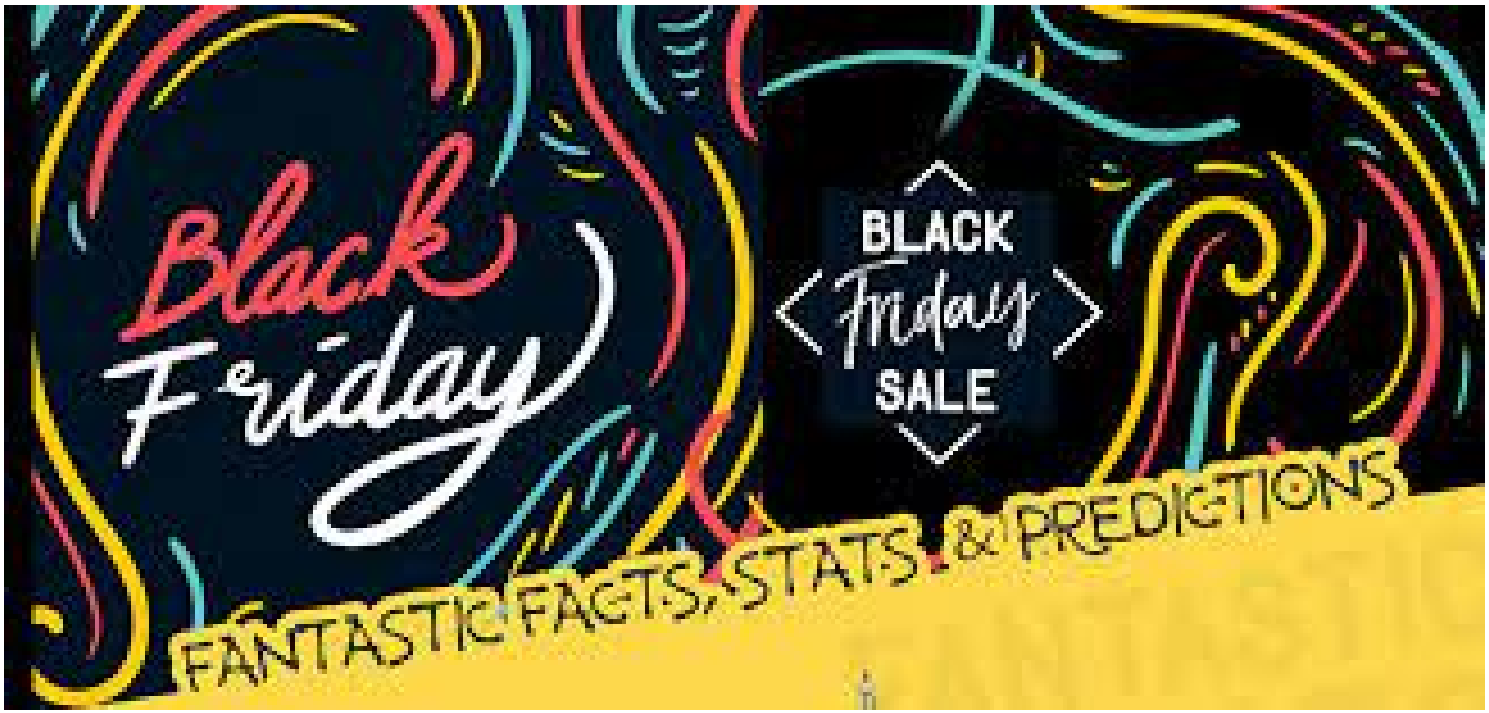
Tran Hoang Quoc Anh
20200044

La Dai Lam
20204918

Luu Trong Nghia
20204888

Table of contents

Introduction



- Black Friday : hottest business day
- Analyze the data from the previous sales and predict the purchase amount

Datasets

Two different data sets: training set and test set.

Training set:

- 550068 rows of data.
- 11 categorical features: **customer profile** and **product detail**.
- *Purchase* is a continuous target label.

User_ID	1000001
Product_ID	P00248942
Gender	F
Age	0-17
Occupation	10
City_Category	A
Stay_In_Current_City_Years	2
Marital_Status	0
Product_Category_1	1
Product_Category_2	6.0
Product_Category_3	14.0
Purchase	15200

Problems

Problem statement

- Given the dataset, the retail company wants to understand customer purchase behaviour against various products given previous month sales
- Create personalized offer for customers against different products

Our goals

- Build a model to predict the purchase amount

SOLUTIONS

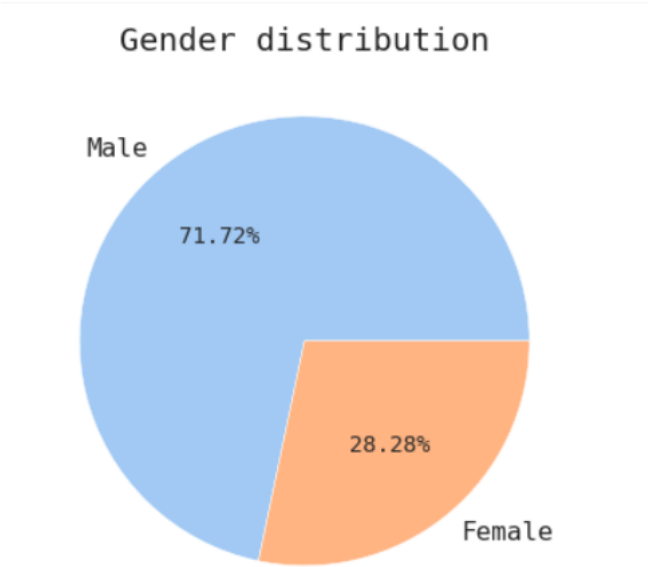
Some tools/libraries

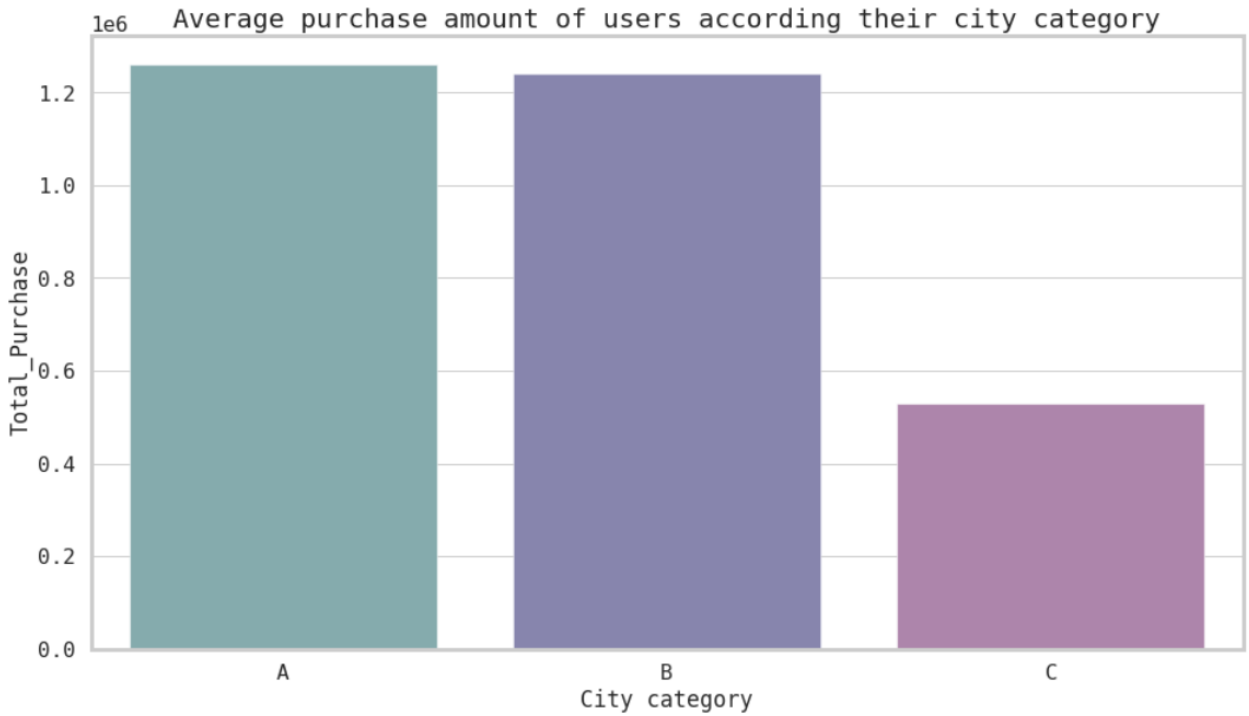
Approach

Our solution follows 4 steps:

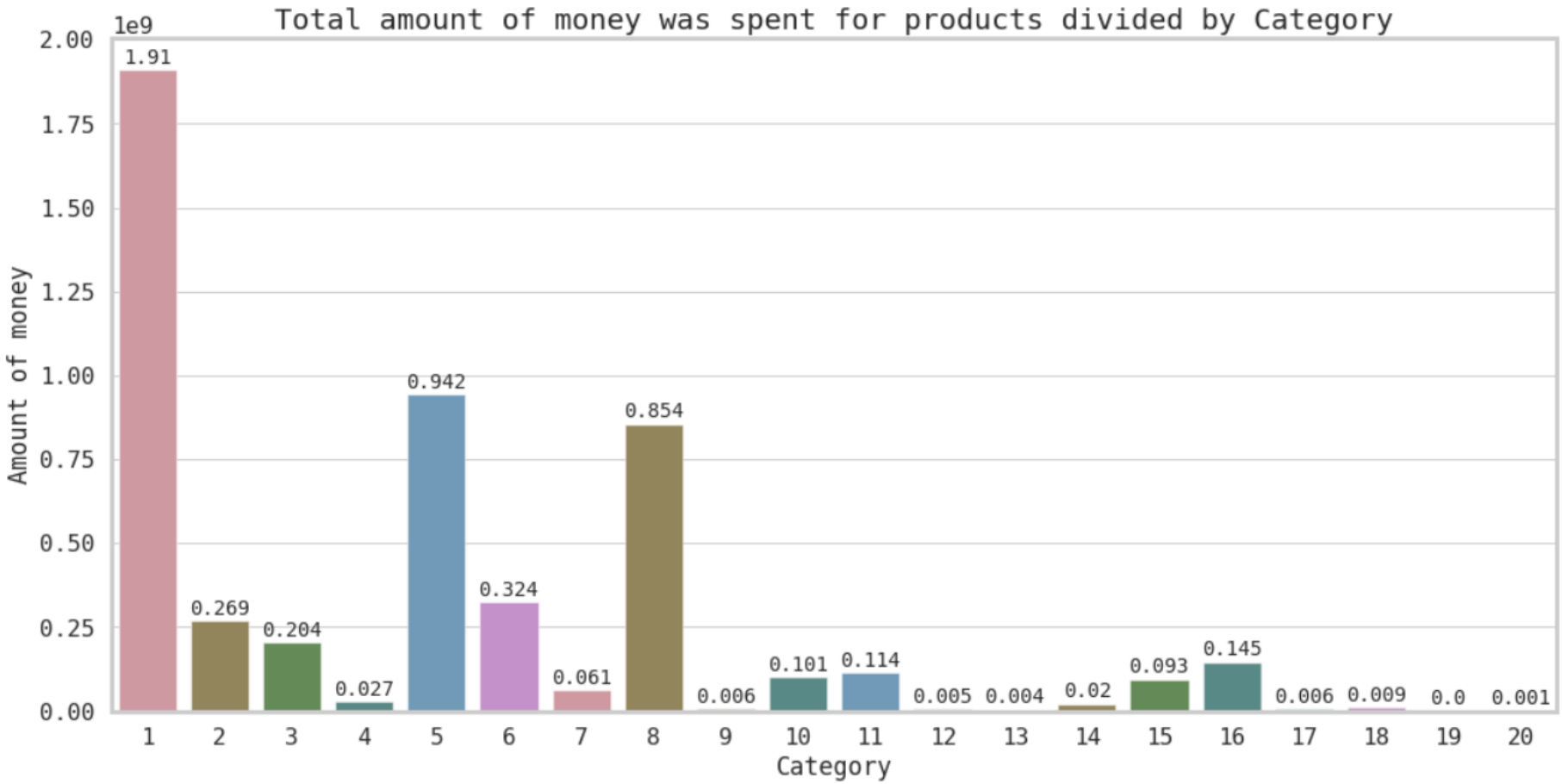
- Explore and find the data patterns
- Choose the metrics to score our model
- Choose a baseline model to set a benchmark
- Compare the results of each model and find the best model

Data Exploration





Data Exploration

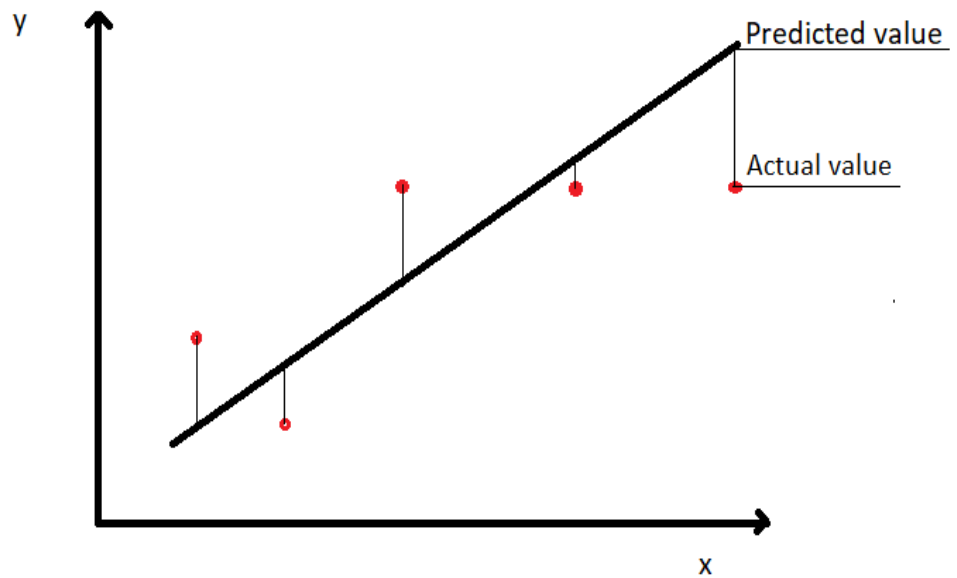


Metrics

Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_i^N ||y_i - \hat{y}_i||^2}{N}}$$

- y_i : i^{th} actual value
- $\hat{y}_i$: i^{th} predicted value
- N : number of instances



The models

- Simple Regression - a baseline model
  - Ordinary Least Square Linear Regression (OLS)
  - Polynomial Regression
- Decision Tree
- Ensemble Learning
  - Random Forest
  - XGBoost
- K-Nearest Neighbors (KNN)
- Collaborative Filtering

RESULTS

Result comparison

Results of different models

	Model	RMSE_mean	Training time(s)	Prediction time(s)	Parameters	Evaluation technique
0	OLS Linear Regression	4614.7	0.172	0.015		5-fold cross-validation
1	Polynomial Regression	4163.05	15.318	0.068	degree = 3	5-fold cross-validation
2	Decision Tree	2733.33	2.879	0.038	max_depth = 15, min_sample_leaf = 20	5-fold cross-validation
3	Random Forest	2738.68	71.483	4.137	n_estimators= 100, min_samples_split= 5, max_features= sqrt	5-fold cross-validation
4	XGBoost	2618.79	84.291	1.143	n_estimators= 200, min_samples_split= 9, subsample = 0.8	5-fold cross-validation
5	K-Nearest Neighbors	2995.14	0.395	1530.0	k = 15	Hold-old (2:1)
6	Collaborative Filtering	4311.77	67.713	338.071		Hold-out (2:1)

Difficulties and limitations

Difficulties

- Do not have common knowledge about the problem
- Do not find the effective way to preprocess the data

Limitations

- Limited hardware resources
- Limited knowledge about theoretical properties of some models (Random Forest)

Conclusion

Summary

- The best model for this problem is XGBoost
- To successful in solving a problem, having great insight about the dataset is crucial
- Finding proper models and suitable set of hyper-parameters for them is really important

Possible extensions

- Understanding the dataset and investigate in more appropriate preprocessing techniques
- Attempting algorithms such as LightGBM and CatBoost

Colab