Introduction

The goal of this project is to analyze Black Friday Sales Data. With this data set, we would like to predict purchase amount shopping for a variety of products. The Black Friday Dataset includes the following information:

- customer demographics: age, gender, marital status, city, length of residence in the city
- product details – ID and product category)
- total purchase amount from last month

In the analysis we will cover the following topics:

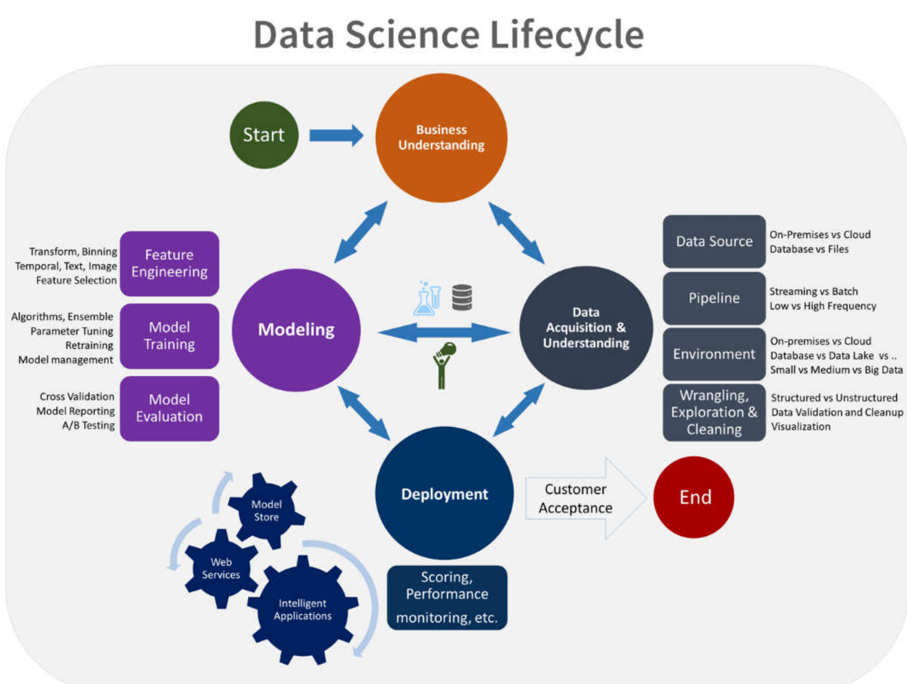
1 - Regression Model Comparison - use several models, analyze stability, hyper parameter tuning and model generalization.

2 - Clustering Algorithm Comparison - We will apply several clustering algorithms and discuss their limitations.

3 - Feature Selection - we will see which features provide the most impact onto the prediction of customer spendings.

Problem Description

We would like to apply key data science lifecycle steps on the specified data in order to determine what kind of relationship we can extract from it. Common business understanding of the sales data is applied,



and since the data has already been partially pre-processed, we start directly with modeling. Lastly, since support and maintenance are out of scope for this project, we deployment step will be skipped.

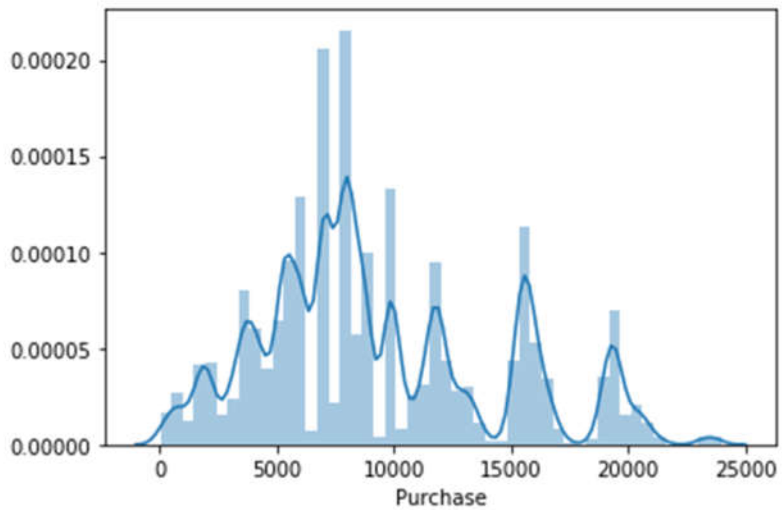
In the given dataset we are provided with 12 columns, most of which require further processing before we can apply ML algorithms.

#	Column	Non-Null Count	Dtype
0	User_ID	550068 non-null	int64
1	Product_ID	550068 non-null	object
2	Gender	550068 non-null	object
3	Age	550068 non-null	object
4	Occupation	550068 non-null	int64
5	City_Category	550068 non-null	object
6	Stay_In_Current_City_Years	550068 non-null	object
7	Marital_Status	550068 non-null	int64
8	Product_Category_1	550068 non-null	int64
9	Product_Category_2	376430 non-null	float64
10	Product_Category_3	166821 non-null	float64
11	Purchase	550068 non-null	int64

Most of the datapoints are available, with the exception of product category 2 and 3. We will treat category 1, as the key category. Optionally, one-hot-encoding can be utilized but this would significantly impact learning times and as we have limited processing power and more effective solution would be to skip this data. Gender, age group, city, stay in city and marital status would need to be one-hot encoded as well. Occupation and product id would be converted into a categorical columns. As the last step we will apply standard scaling to the data set. Processed data set contains 18 columns:

User_ID	550068 non-null	int64
Product_ID	550068 non-null	int32
Gender	550068 non-null	uint8
Occupation	550068 non-null	category
Marital_Status	550068 non-null	int64
Product_Category_1	550068 non-null	int64
Age_18-25	550068 non-null	uint8
Age_26-35	550068 non-null	uint8
Age_36-45	550068 non-null	uint8
Age_46-50	550068 non-null	uint8
Age_51-55	550068 non-null	uint8
Age_55+	550068 non-null	uint8
City_Category_B	550068 non-null	uint8
City_Category_C	550068 non-null	uint8
Stay_In_Current_City_Years_1	550068 non-null	uint8
Stay_In_Current_City_Years_2	550068 non-null	uint8
Stay_In_Current_City_Years_3	550068 non-null	uint8
Stay_In_Current_City_Years_4+	550068 non-null	uint8

Looking at the purchase amounts, we see that most purchases fall within 5,000 – 10,000 range, and around 15,000, and 20,000 ranges.



We can observe non-uniform distribution of values in this dataset, and as the result our models might be biased towards high-frequency values, we should be aware of this issue. As we are r

Solution Description

In order to predict spending amounts (regression problem), we trained key machine learning models, specifically:

Linear models – simple linear, polynomial, ridge and lasso
Decision trees – decision trees, random forest, gradient boost, XGBoost
Support vector machine – SCV with radial basis function kernel
KNN – K-neighbours regressor
ANN – Multi-layer perceptron network

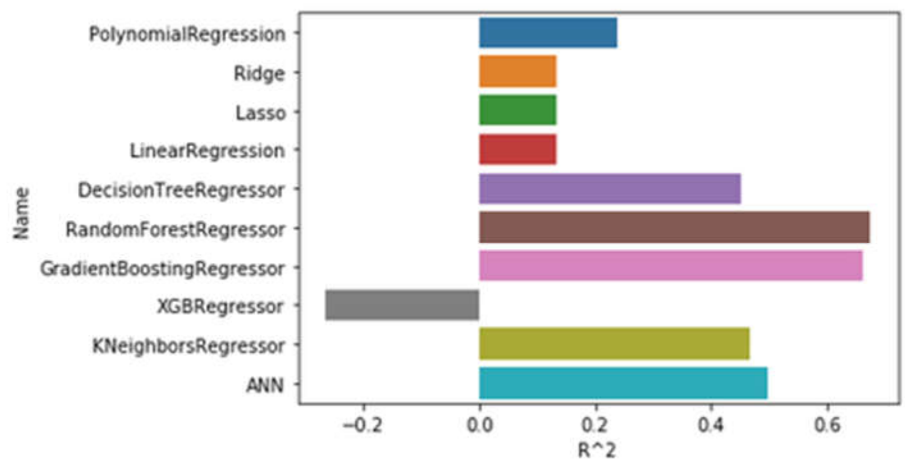
For measuring results the following metrics were collected for each of the learning model:

Mean absolute error - difference between actual and predicted value
Mean squared error – Squares the difference between actual and predicted value
Root-mean squared error - root of the MSE
R^2 – Compares model with a constant baseline

During the explorative analysis default parameters were used for most of the regressors. Further after running regression metrics were stored in an array for comparison.

Results

The diagram below shows the summary of R2 metrics. Top algorithms for this solution are Random Forest, Gradient Boosting and the neural networks. Random Forest turned out to be the best predicting algorithm with RMSE metric of 2866, absolute error of 2105 and R2 value of 68%.



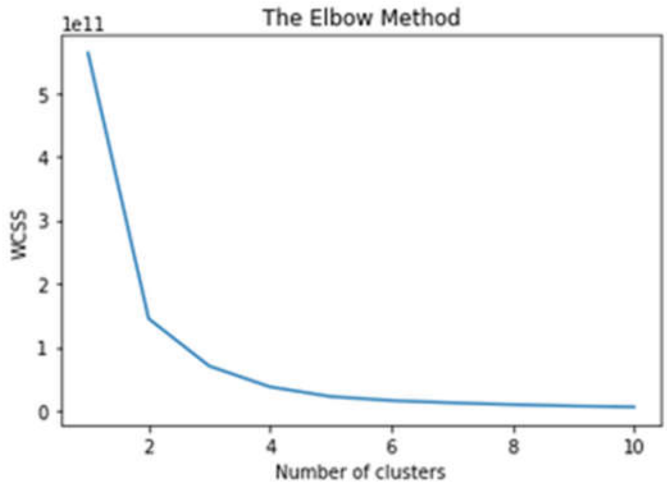
Detailed performance metrics of the algorithms are listed in the table below

Regressor	MAE	MSE	RMSE	R^2
RandomForestRegressor	2104.10	8211631.21	2865.59	67.54%
GradientBoostingRegressor	2214.39	8516361.50	2918.28	66.33%
ANN	2280.81	9094922.78	3015.78	64.04%
KNeighborsRegressor	2632.36	13454954.47	3668.10	46.81%
DecisionTreeRegressor	2630.83	13872400.00	3724.57	45.16%
PolynomialRegression	3309.96	19261760.00	4388.82	23.85%
Ridge	3593.79	21952040.00	4685.30	13.21%
LinearRegression	3593.79	21952040.00	4685.30	13.21%
Lasso	3593.78	21952040.00	4685.30	13.21%
XGBRegressor	4221.34	32038236.71	5660.23	-26.66%

Feature selection algorithms did not yield significant improvement of performance metrics of the key algorithms.

SVN algorithms were not effective due to slow ($O(N^2)$) run-time performance. It was unable to converge within a reasonable timeframe for the purposes of the assignment.

K-Means algorithm identified 5 segments within the dataset using the Elbow method:



Conclusion

Analysis of the Black Friday Sales Data provided decision tree and neural network algorithms as clear winners among supervised learning algorithms. These algorithms require further investigation and optimization in order to obtain the best prediction results. Further elaboration on feature engineering may also impact the predictions. Application of K-Means algorithm also identified five key segments that can be used for categorizing the data.