1. 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)

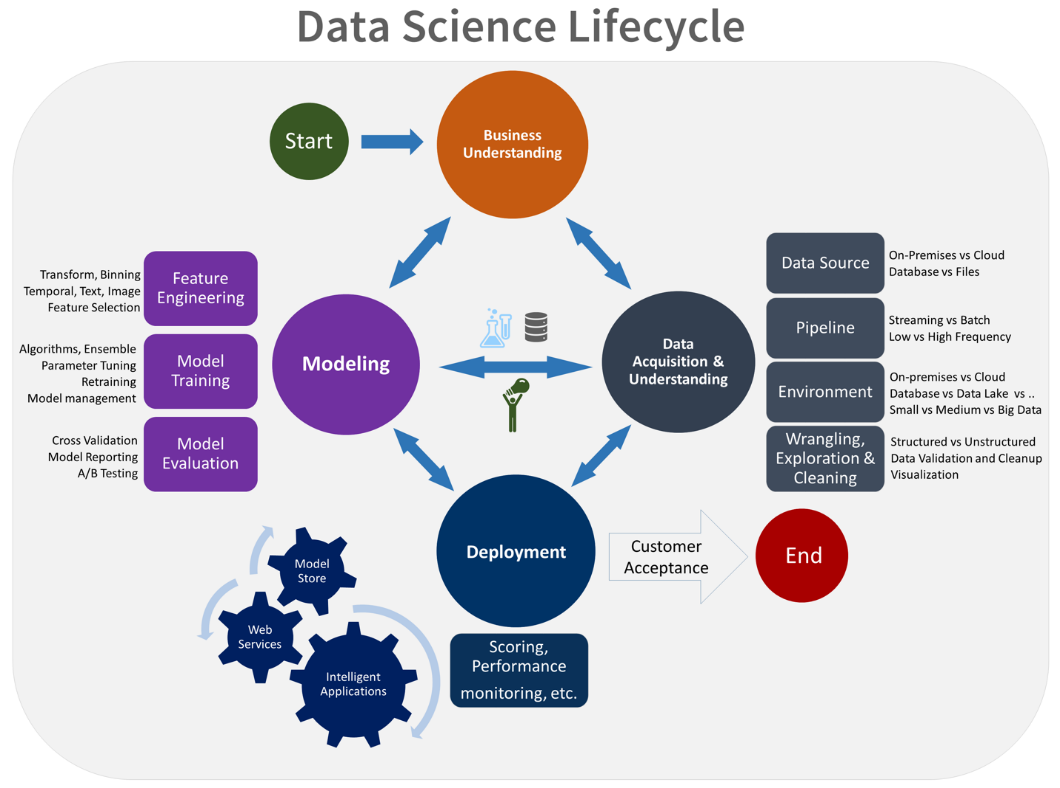
- total purchase amount from last month

In the analysis we will cover the following topics:

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

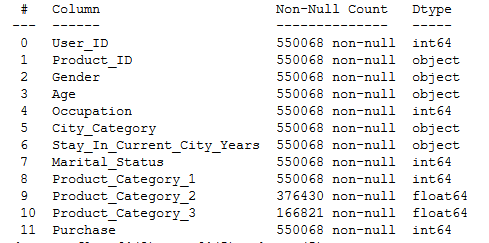
2 - Clustering Algorithm Comparison - We will apply several clustering algorithms and discuss their limitations.

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

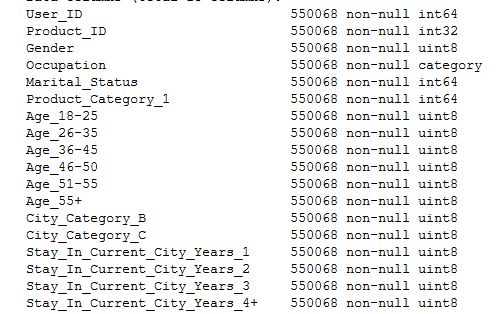
1. Problem Description

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.

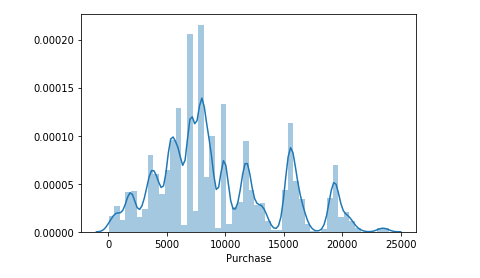
In the given dataset we are provided with 12 columns, most of which require further processing before we can apply ML algorithms.



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:



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.

1. Solution Description

In order to predict spending amounts the fo

1. Results

Dd

1. Analysis

D

1. Conclusion