**MEF UNIVERSITY**

# Predicting the Impact of Product and User Features on the Sales in an E-Commerce Site

**Capstone Project**

**Samet Boyacı**

**İSTANBUL, 2018**

**MEF UNIVERSITY**

# Predicting the Impact of Product and User Features on the Sales of Product in an E-Commerce Site

**Capstone Project**

**Samet Boyacı**

**Advisor: Asst. Prof. Dr. İrem Zeynep Yıldırım**

**İSTANBUL, 2018**

# MEF UNIVERSITY

Name of the project: Sales Predicting the Impact of Product and User Features on the Sales of Product in an E-commerce Site
Name/Last Name of the Student: Samet Boyacı

I hereby state that the graduation project prepared by Samet Boyacı has been completed under my supervision. I accept this work as a "Graduation Project".

12/09/2018
Asst. Prof. Dr. İrem Zeynep Yıldırım

I hereby state that I have examined this graduation project by Samet Boyacı which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

12/09/2018
Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Samet Boyacı and agree that the student has satisfied all requirements.

**THE EXAMINATION COMMITTEE**

| Committee Member | Signature |
|---|---|
| 1. İrem Zeynep Yıldırım | …….. |
| 2. …….. | …….. |

## Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

---

Name Date Signature

# EXECUTIVE SUMMARY

PREDICTING THE IMPACT OF PRODUCT AND USER FEATURES ON THE SALES OF PRODUCT IN AN E-COMMERCE SITE

Samet Boyacı

Advisor: Asst. Prof. Dr. İrem Zeynep Yıldırım

AUGUST, 2018, 31 pages

In recent years, the ratio of online shopping to total shopping has been increasing continuously. Many factors affect sales of e-commerce sites. Prior to purchasing, users are concerned whether the features of the products they are interested in match their own needs or not. In this study, the most important factors in the sales of the products which are the features of the products, attributes of the sellers and interactions with product investigated.

A model was developed based on available fashion products in a market where individual users could sell second-hand textiles and accessories. Using this model, we tried to predict which products would be sold by examining at the features of the products, attributes of sellers and interactions with the product.

Different algorithms were investigated for predicting sales and the results were reported. By comparing the outputs, the most successful algorithm and the most important features affecting sales were identified.

As a result of this study, it was determined that the most efficient algorithm was the decision tree model. When the inputs of the model were examined, it was determined that the most important features affecting salability were the interactions with the products such as the number of likes and bids.

**Key Words**: E-Commerce, sellability, selling of product, features of product, classification

# ÖZET


E-TİRCARET SİTESİNDE ÜRÜN, SATICI ve SİTE ÖZELLİKLERİNİN ÜRÜN SATIŞINA OLAN ETKİSİNİN TAHMİN EDİLMESİ

Samet Boyacı

Tez Danışmanı: Dr. İrem Zeynep Yıldırım

AĞUSTOS, 2018, 31 sayfa

Son yıllarda internet üzerinde yapılan alışverişlerin toplam alışverişe oranı her geçen gün artmaktadır. İnternet üzerinden satış yapan e-ticaret sitelerinde satışı etkileyen bir çok faktör bulunmaktadır. Kullanıcılar satın almadan önce ilgilendikleri ürünlerin özelliklerinin kendi ihtiyaçlarını ne kadar karışıladıkları ile ilgililenmektedir. Bu çalışmada ürünlerin satışında en önemli etkenlerden biri olan ürünlerin özelliklerinin ve o ürünün sitede yaratttığı etkileşimin satışa etkisi araştırıldı.

Bu model sayesinde ürünlerin özellikleri ve ürün ile olan etkileşimine bakarak hangi ürünlerin satılabilir olduğunu tahminlenmeye çalışıldı.

Farklı algoritmalar kullanarak ürünlerin satılabilirliği tahmin edildi ve sonuçları değerlendirildi. Çıktılar birbirleri ile karşılaştırıldı ve en başarılı yöntem belirlendi. Ürünün satılabilirliğini etkileyen en önemli özellikler belirlendi.

Çalışma sonucunda problemin çözümü için en verimli modelin "Karar Ağacı Algoritması" olduğu tespit edildi. Modelde kullanılan girdiler incelendiğinde satılabilirliği etkileyen en önemli özelliklerin beğeni sayısı, teklif adedi gibi ürünlerin sitedeki etkileşimleri olduğu tespit edildi.



**Anahtar Kelimeler**: Ürün özellikleri, satılılabilirlik, sınıflandırma, e-ticaret

# TABLE OF CONTENTS

# LIST OF FIGURES



# LIST OF TABLES

# 1. INTRODUCTION

At the beginning of the study, this section gives important information about the e-commerce sector. Thus, the problem can be identified easily by understanding the business model better. After that, you can find information about the website which is obtained dataset for this study. Especially this section explains the structure and features of the site.

Along with the recent development of e-commerce, known commercials start to change around the world. Based on recent research, e-commerce market share in 2018, as a percentage of all retail sales, is expected to increase to 11.9% - up from 3.5% a decade ago (Lui, 2018). Also, this share is expected to increase every year. Figure 1 shows time versus e-commerce share of global retail.

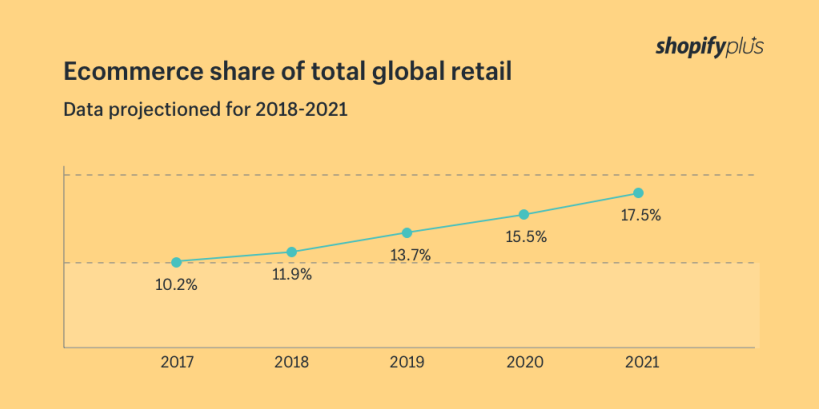



**Figure 1. E-commerce share of total global retail (Lui, 2018)**

Due to the young population in Turkey, e-commerce is growing much faster than in other countries. Turks who love to experience innovations have also made rapid adaptation to e-commerce (Tekel, 2014).

Trust is one of the most important factors for users to use e-commerce sites. If users are satisfied while buying process, the buying frequency increases and they spread this confidence to their surroundings.

Online shopping costumers do not need to appear as a person in physical stores, so they will not be interfered by various outside factors and can go shopping according to their own thoughts, which leads their purchase behaviors more complex and random. However, the purchase behaviors of online customers can be quantified as concrete data such as the amount of click, page visits, orders and purchase. All the data have been obtained easily and accurately thanks to recording (logging) structure of the website. Using the historical data of

the products on the site, this project aimed to discover the relationship between purchasing habits and product features.

### 1.1. Secondhand Clothing and Accessories Sites

With the globalization of the world and the increase in purchasing power, shopping habits change day by day. Buyers no longer consider buying high-priced products. For this reason, many of the products stay in our homes even though they are not used at all.

Some e-commerce sites build new business model to provide sellers who want to sell products. This product's condition are generally on labels as "new" or "very little used". Sellers sell the products lower than buy prices at shops on these sites.

In the second hand clothing site, there are two different user roles: buyer and seller. Sellers take pictures of the products they want to sell and upload them to the site. Also they provide detailed information about the product. For example, a user who wants to sell a watch that he/she rarely uses. Users enter information about the brand, model, color and price on the site.

Buyers can buy the products they are interested in from these sites. These sites can provide details about the products or brands they are interested in. Buyers can also purchase the product by contacting the seller of the product.

After the sale is made, the site will generate revenue by receiving commission on the sale price. As sales on the site increase and products with higher prices are sold, the income of the site increases. In other words, the business model of the site is based on the income from commissions on sales.

In an e-commerce site, that sales of second-hand clothes, the life cycle of the products continues by following steps:

Buyers and sellers create membership on the site.

Sellers upload the products they want to sell on the site with their photos.

Buyers can bid or directly purchase for the products they are interested in.

The receivers are directed to the checkout screen and payment is made by credit card.

Payment is held in the temporary pool of the site.

Seller sends the product to the buyer by cargo.

After buyer checks and accepts the product, buyer confirms the product on website.

The money in the pool is deposited to account of the seller after the commission of the site is cut.

### 1.1.1. Point and Comment System

In these types of sites, buyers are able to score buyers' experience after purchasing. The confidence levels of the sellers are based on this score.

Buyers follow the products sold by high-ranking sellers more. For this reason, every vendor is obliged to provide correct information to the buyer when selling the product.

These sites also has comment systems, buyers may make comments for each product. These comments may contain details about the products. Buyers may request additional information about the product from the sellers. Too much interest in the product is an indication that the product is attracting attention from many buyers.

At the same time, buyers can report vendors who add misleading information about their products. The site launches a review of products that are reported by a user on a certain number. If the product is published with misleading information, the product is closed by the site. A warning is issued to the seller. Sellers who receive similar warnings are permanently removed from the site.

These two systems were created to ensure that both sellers and buyers can shop safely.

### 1.1.2. Purchasing System

Seller upload products to website for selling. When sellers upload products, they choose purchasing type. Purchasing type can be direct sales or bidding. Direct sales mean buyers can only buy with the price defined by the seller at the beginning. Buying price doesn't change unless sellers want to change price. If buyers like the product, he/she can buy with actual price.

In a bidding process, buyers can bid for a price lower than the sale price of the products they want to buy. Website allow only bids that is lower than actual price.

It is up to the seller to accept or reject the offer of the buyer. The seller may reject the offer. If the seller accepts the bid offered by the buyer, the sale is made at the proposed lower price. Buyers cannot see the bid price from other buyers.

Otherwise, the seller may counter bid against the bid of the buyer. In this scenario, the same situation applies to the buyer. Both users can bid until the bid is accepted or rejected.

With the bidding system, even if the current price is not suitable, alternative prices are available to buy products.

Products that receive more bids than other products, always have high sellability. Because buyers bids show us these products have a high demand.

# 2. PROJECT DEFINITION

This section explains the purpose and scope of the project. In this way, more detailed information will be provided about the point to be reached.

## 2.1. Problem Statement

The main problem was that it was not known whether new products added to the website are salable or not. Before this study, products were featured on the main page of the website according to some assumptions. The main goal is to identify the products that can be sold with self-learning intelligent models in order to increase sales on the website.

Products similar to previously sold products should be labeled as salable. In this manner, CRM teams will be able to offer these products both on the main page of the site and on the banners.

## 2.2. Project Objectives

The objective of the project is to automatically identify products similar to those previously sold under certain constraints. In this way, more sales will be made on the site and income will be increased.

The breakdown of project objectives are as follows:

Analyzing the dataset provided by Secondhand clothing website,

Analyzing and extracting meaningful information in the data,

Identifying the most suitable models and parameters,

Running the model to label new products,.

Measuring success of model and the algorithm,

Comparison of outputs.

## 2.3. Project Scope

The scope to be followed while working on the project will consist of the following items:

Converting the data into a workable tabular form,

Cleaning missing values,

Removing irrelevant features,

Creating new features with feature engineering,

Data analyzing with exploratory data analysis,

Splitting test and train data,

Training models,

Evaluation of model results,

Comparison of model results,

Choosing the best model to solve the problem.

# 3. ABOUT THE DATA

## 3.1. General Description

The dataset includes features of products on the e-commerce site that sells second-handed clothes and accessories. This data was chosen because it contains the relevant product features which affect sellability.

Data is the information that has been translated into a form that is efficient for movement or processing. Data transforms to information after analysis phases. Columns of data would be the input you have fed to the system. For example, size, location, number of rooms are features in house pricing data. Labels are expected result of data the analysis. They would be the output you are expecting. In this example, price of house is label.

The data will cover the products uploaded to the website between 1 June 2018 – 1 July 2018. Products that has approved, sold and closed status will be included in the scope of the study. Products that are rejected or pending approval by the editor will not be included in the project.

The total number of rows were 729,141 in the beginning. Before setting up the data model, the dirty data should be removed before the analysis. The data that are unrelated for the research was removed from the data set. For example, some of the products which are rejected by editors or do not have critical features are removed. These rules will be explained in detail in the pre-processing section.

**Features:** Features are individual independent variables that act as the input in your system. Prediction models use features to make predictions. New features can also be obtained from old features using a method known as 'feature engineering'. More simply, you can consider one column of your data set to be one feature. Sometimes these are also called attributes and the number of features are called dimensions.

The dataset involves all the possible information available for each product. It includes total of 34 columns or potential features to be used for the analysis. The dataset contains various related to the product such as date of publication, brand, model, price etc. These areas are usually categorical data.

In addition, the model will be included extracted features from the data set other than product attributes. These features usually consists of numeric variables. For example, the

number of products' comments. These features will be explained in detail in the future extraction section.

**Labels:** Labels are the final output. You can also consider the output classes to be the labels. When data scientists speak of labeled data, they mean groups of samples that have been tagged to one or more labels.

As a result, it appears appropriate to the classification algorithms due to the large number of features and labels contained in the data set. There will be some constrains in the model to deliver the business needs. The average sales of the products uploaded to the site in a week will be considered as sales. The products sold for a longer than a period of one week will not be taken into account. Therefore, selling the products within a certain period of time to solve the problem will be considered as a success factor.

### 3.2. Features of Dataset

This section explains dataset with four heading.

1. **Feature Group:** Three main feature groups are present in the dataset. First group is related to the product. Product features describe the product detail. In the second group, there are features related to the seller. Last group is related to the measures. Dataset has calculations like interactions of product and seller rate.
2. **Feature Names:** A name gives information about the contents of that field.
3. **Description:** The details about the field are explained in the title.
4. **Data Type:** Type of related field. Data types are a guideline for analyzes to be performed on the field. Numerical fields are measurable fields that can perform various mathematical calculations. Date fields are fields that represent a specific time or date. Categorical fields consist of characters that take a certain group.

Table 1 shows the list, explanations and data types of the features in the dataset used:

**Table 1. Features of dataset**

| Group | Feature Names | Description | Data Type |
|---|---|---|---|
| 1 | Product Id | Unique Product ID, Every product has unique identifier. | Numerical |
| 1 | Create Date | Product creation date on website | Date |
| 1 | Product Status | Last product status. Product status can ben change by times. | Categorical |
| 1 | Price | Sales price of product. Price should be greater than zero. | Numerical |
| 1 | Original Price | Original price of product in other stores. | Numerical |
| 1 | Allow Bidding | Is bidding allowed? | Categorical |
| 1 | Condition | Condition of product. | Categorical |
| 1 | Sub Category Name | Sub category names. | Categorical |
| 1 | Sub Category Group | Sub category group. | Categorical |
| 1 | Parent Category Name | Parent category names. | Categorical |
| 1 | Parent Category Group | Parent category group. | Categorical |
| 1 | Brand Name | Brand name. | Categorical |
| 1 | Brand Type | There are different classes for each product. | Categorical |
| 1 | For Kids Baby | Kids or baby brand flag. | Categorical |
| 1 | For Woman | Woman brand flag. | Categorical |
| 1 | Size | Size of product. | Categorical |
| 1 | Size Section | Every product has a different size section. | Categorical |
| 1 | Color | Product color. | Categorical |
| 1 | Sales Date | Sales date of product. | Date |
| 1 | Sales Label | Is product sold? | Categorical |
| 2 | Membership Date | Membership date of seller | Date |
| 2 | Seller Trust Level | Brand class defined by site according to seller rating. | Categorical |
| 2 | Seller Feedback Average | Feedback average points for seller users. Range: 0-5 | Numerical |
| 2 | Nof Follower For Seller | Number of follower for seller. | Numerical |
| 3 | Nof Comment | Number of comment. | Numerical |
| 3 | Nof Likes | Number of likes. | Numerical |
| 3 | Nof My Brand By User | Number of users which add "my brand" for product brand. | Numerical |
| 3 | Nof My Size By User | Number of users who add "my size" for product size. | Numerical |
| 3 | Nof Bid | Number of bids. | Numerical |

### 3.3. Model and Data Constraints

Some constraints about the project and the data can be listed as follows:

The scope of the project is about only estimating the published products that are targeted.

The label of the product must be created in the system before the product was published on the website.

Model should be re-calculable after getting product-related interactions from the website.

After developing the model, only 10% of the customers will be released for the first part. At the end of tests, it will be published to all customers if it is success.

While the model was designed, it should be noted that new features might be added in the future.

# 4. METHODOLOGY

Methodology section explains various approaches to solve the problem. First, the data will be tried to be understood by explanatory data analysis methods. Then, a statistical model will be developed to solve this problem with machine learning approaches.

## 4.1. Data Preprocess

Initially, the preparation phase must be passed before the data was processed. Relational data should be integrated into different sources.

For this, (Structured Query Language) SQL statements were written at the database level to merge the data into different tables. To understand which tables combine with each other, the data model of the tables needs to be crated. Figure 2 shows the relationships between the tables in the source database.

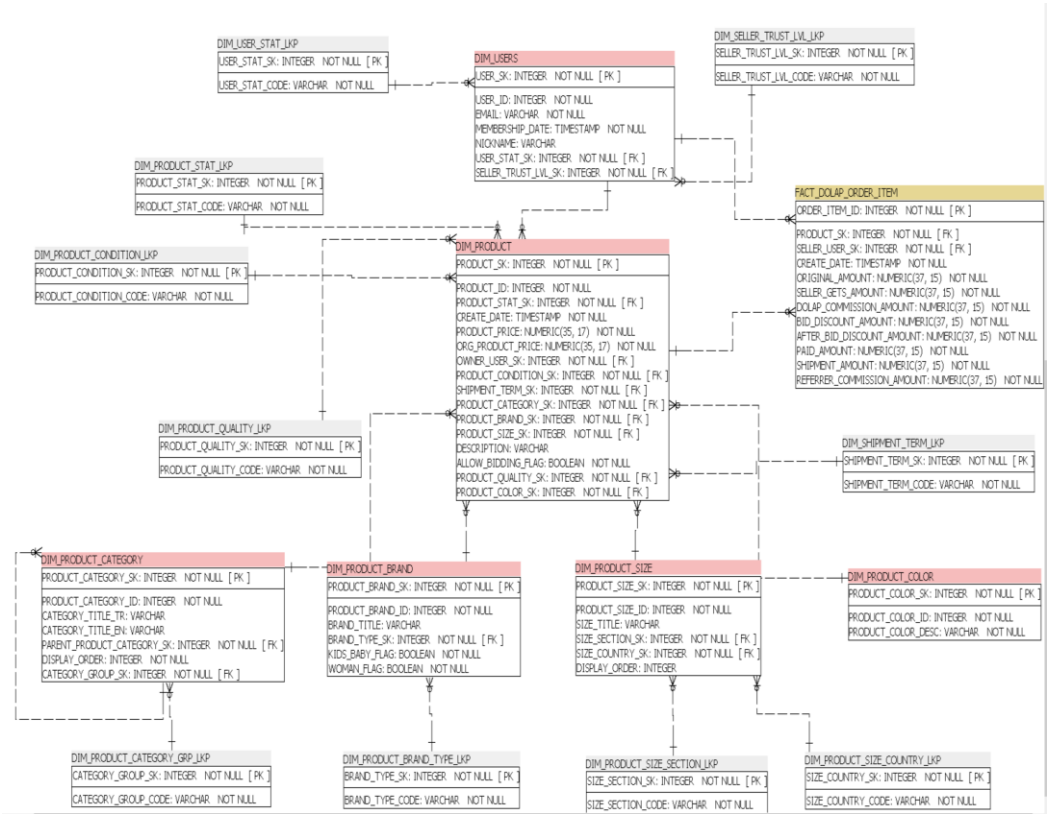



**Figure 2. Data model of source tables**

### 4.2. Data Cleaning

This section explains handling the missing data in the dataset. In addition to data visualization and analysis of correlation and adjusted R-square scores, the ratio of missing values in each feature was analyzed as well.

Since the precision of the rule-based approach is important, imputing the missing values was not preferred. The algorithm could still catch similarities when same features of both entities have missing values. That being said, selecting the features with the great number of missing values would not be advantageous to detect the similarities.

There are two options for handling the missing values. First, one is to delete related rows that has missing value. This can lead to decrease number of rows. Another way is to fill missing values. There are different approaches filling values like mean, median, min or max values (Gelman and Hill, 2007).

### 4.3. Exploratory Data Analysis

During the exploratory data analysis phase, the impact of features on sellability was measured through descriptive statistics. This section examines the correlation between features.

The main motivation of this phase is to simplify the model by reducing the number of features by selecting the ones that are the most significant. To do so, the features with the least missing values are targeted.

### 4.4. Model Building

In this project, various classification techniques are used. In machine learning, classification techniques are worked through the processes that appear in Figure 3.

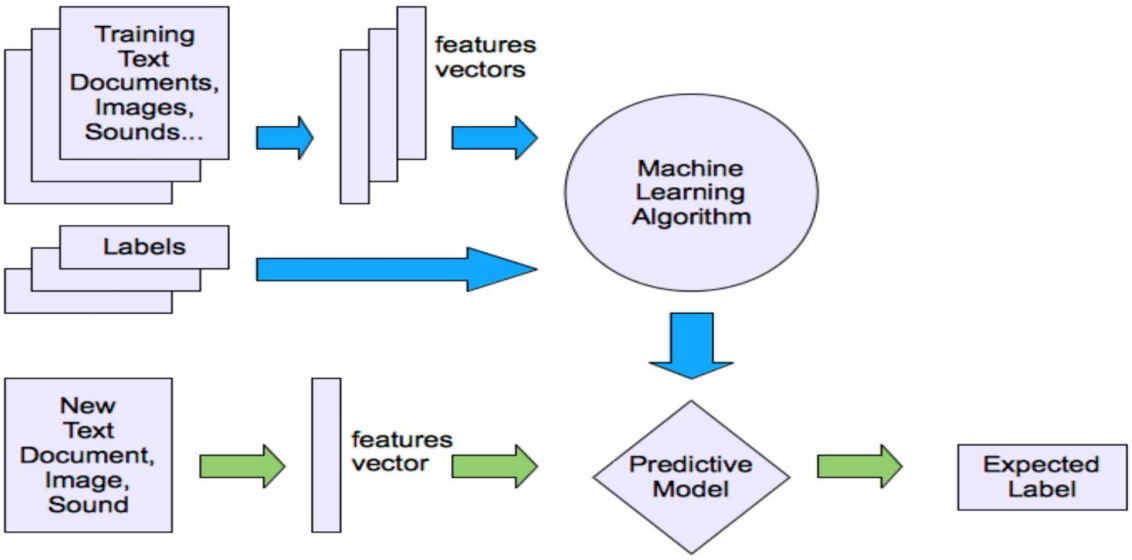



**Figure 3. Steps of supervised learning (Kumar, 2018)**

Firstly, data that has features and labels trained with classification algorithms. Sometimes data cannot be stored in vector or table form. It should be unstructured data. Unstructured data should be transformed to structured data. After that, new data labels are estimated with this learned model. In Figure 3, you can see the steps of supervised learning a machine in a stream.

In this project, models are built with following algorithms:

**Decision Trees:** Decision trees are the most commonly used because of its ease of implementation and ease in understanding as compared to other classification algorithms. Decision Tree classification algorithms can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm. (Anyanwu and Shiva, 2009).

**Support Vector Machine:** The SVM algorithm tries to determine the classes in the data with decision boundaries. Observations on different sides of the decision boundaries are in different classes. The best decision boundaries separates data with the largest margin between two classes and the distance from it to the nearest data point on each side is maximum. Academic studies in e-commerce sales data show that SVM models trained with the common Gaussian kernel are overfitting the data and do not generalize well. (Liu and Li, 2016). Figure 4 shows the example display of different SVM kernels.

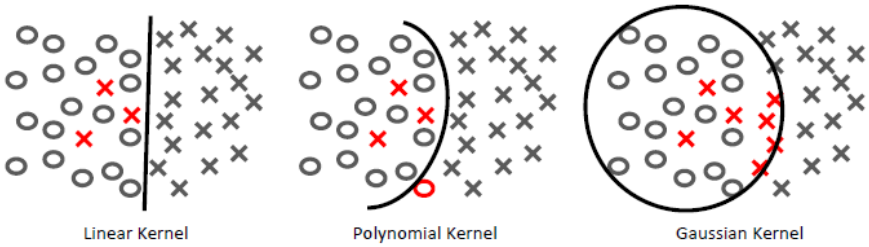



**Figure 4. SVM decision boundary (Chen, 2014)**

**XGBoost:** XGBoost is a short for extreme gradient boosting. It is a library designed and optimized for boosted tree algorithms. It's main goal is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate for large scale tree boosting.

Most important feature of XGBoost is speed. Gradient boosted trees, have to be built in series so that a step of gradient descent can be taken in order to minimize a loss function. (Steinweg-Woods, 2016)

### 4.5. Model Performance Metrics

This section explains understanding result of models. There are a lot of model metrics. Each metrics use different cases. It depends on type of model and problem.

**Confusion Matrix:** For a binary classification problem the table has 2 rows and 2 columns. Across the top is the observed class labels and down the side are the predicted class labels. Each cell contains the number of predictions made by the classifier that fall into that cell. Table 2 shows detail of the Confusion Matrix.

**Table 2. Confusion matrix**

| | Predicted Class | | |
|---|---|---|---|
| Actual Class | | Yes | No |
| | Yes | True Positive | False Positive |
| | No | False Positive | True Negative |

**True Positives (TP):** These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. For example, if actual class value indicates that this passenger survived and predicted class tells you the same thing.

**True Negatives (TN):** These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. For example, if actual class says this passenger did not survive and predicted class tells you the same thing.

**False Positives (FP):** When actual class is no and predicted class is yes. For example, if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

**False Negatives (FN):** When actual class is yes but predicted class in no. For example, if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

**Accuracy Score:** Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

**Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate (Joshi, 2016).

# 5. RESULTS

## 5.1. Results of Exploratory Data Analysis

Key implications such as average, minimum/maximum values and number of missing values are important for discovering the data. The basic statistical information about the data set can be found in Table 3.

**Table 3. Descriptive statistics of data**

| Feature | Unique Value Count | Missing Value Count | Min | Max | Mean |
|---|---|---|---|---|---|
| product_create_date | 30 | 0 | | | |
| product_close_date | 54 | 601138 | | | |
| product_status | 3 | 0 | | | |
| price | 853 | 0 | 20 | 50000 | 70,940 |
| original_price | 1419 | 200322 | 20 | 50000 | 197,200 |
| condition | 3 | 0 | | | |
| product_quality | 4 | 0 | | | |
| sub_category_name | 101 | 0 | | | |
| sub_category_group | 2 | 0 | | | |
| parent_category_name | 9 | 0 | | | |
| parent_category_group | 2 | 0 | | | |
| brand_name | 1789 | 0 | | | |
| brand_type | 4 | 0 | | | |
| for_kids_baby | 2 | 0 | 0 | 1 | |
| for_woman | 2 | 0 | 0 | 1 | |
| size | 75 | 0 | | | |
| size_section | 8 | 0 | | | |
| colour | 39 | 0 | | | |
| seller_membership_date | 696 | 0 | | | |
| seller_trust_level | 6 | 0 | | | |
| seller_feedback_average | 6 | 0 | 0 | 5 | 3,170 |
| nof_reporter | 9 | 0 | 0 | 8 | 0,005 |
| average_price_of_brand | 5658 | 0 | 0 | 43465 | 972,170 |
| nof_comment | 107 | 0 | 0 | 206 | 1,250 |
| nof_follower_for_seller | 2576 | 0 | 0 | 20115 | 418,300 |
| nof_likes | 137 | 0 | 0 | 259 | 1,870 |
| seller_nof_bloked_by_others | 50 | 0 | 0 | 190 | 1,520 |
| nof_my_brand_by_user | 773 | 0 | 3 | 63837 | |
| nof_my_size_by_user | 104 | 0 | 0 | 76372 | |
| nof_bid | 41 | 0 | 0 | 52 | 0,299 |
| sales_label | 2 | 0 | 0 | 1 | 0,071 |

### 5.1.1. Brand

Sellers select brands when they upload the product on website. Seller should choose the correct brand name for products. If sellers choose wrong name of brand, editors can reject before it is published on the website.

Low-known brands might not be included on the website. If the brand is not available on website brand list, sellers select ‘other’ option. Figure 5 shows distribution of brands in dataset.

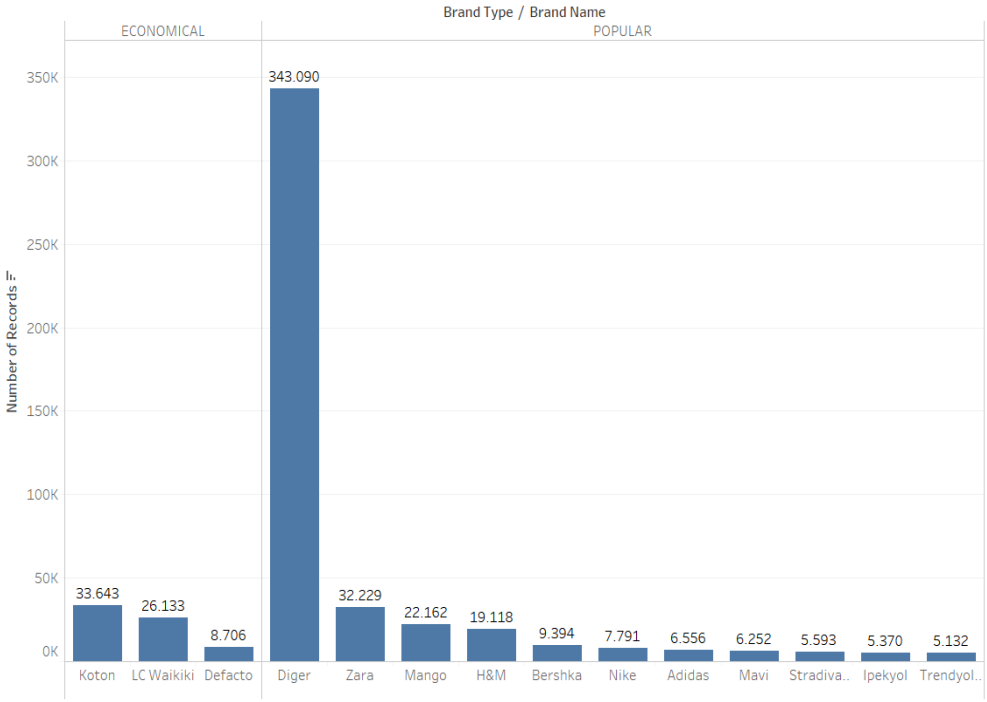



**Figure 5. Frequency of brands**

As seen in Figure 5, Koton and LCW are the brands in economic category, which has the most number of products on the website. Also Zara, Mango and H&M has important part in Popular category. Another points in figure, almost 40% of the product brand is selected as “other”. Especially accessories and top wear categories has a lot of “other” brand product. Figure 6 show which types of brands include in dataset.

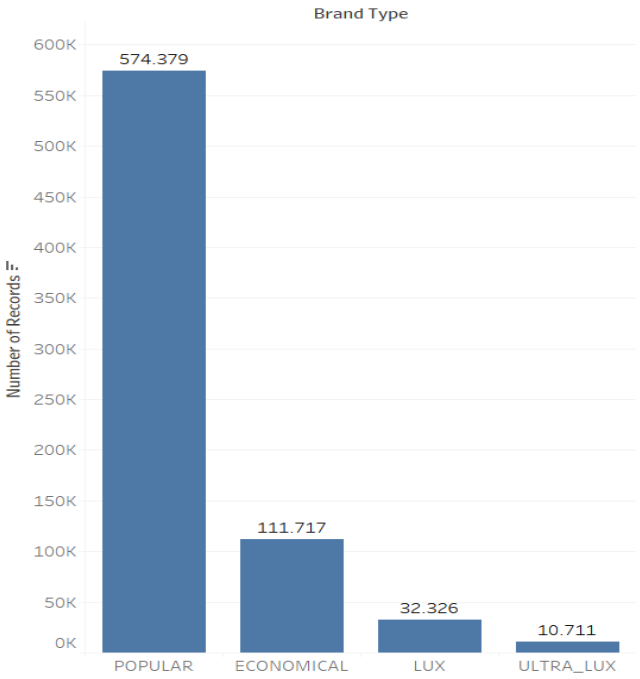



**Figure 6. Frequency of brand type**

As seen in Figure 6, 78% of products have a popular category. There are few product in Lux and Ultra Lux brands. That means, sellers tend to sell popular brands more than luxury brands.

### 5.1.2. Category

Sellers choose parent categories and sub-categories when they upload products on the website. The locations of the products within the site are made according to these selected categories. The editors can change the product category after the seller have entered the categories incorrectly. Category-based distributions of products are shown in Figure 7.

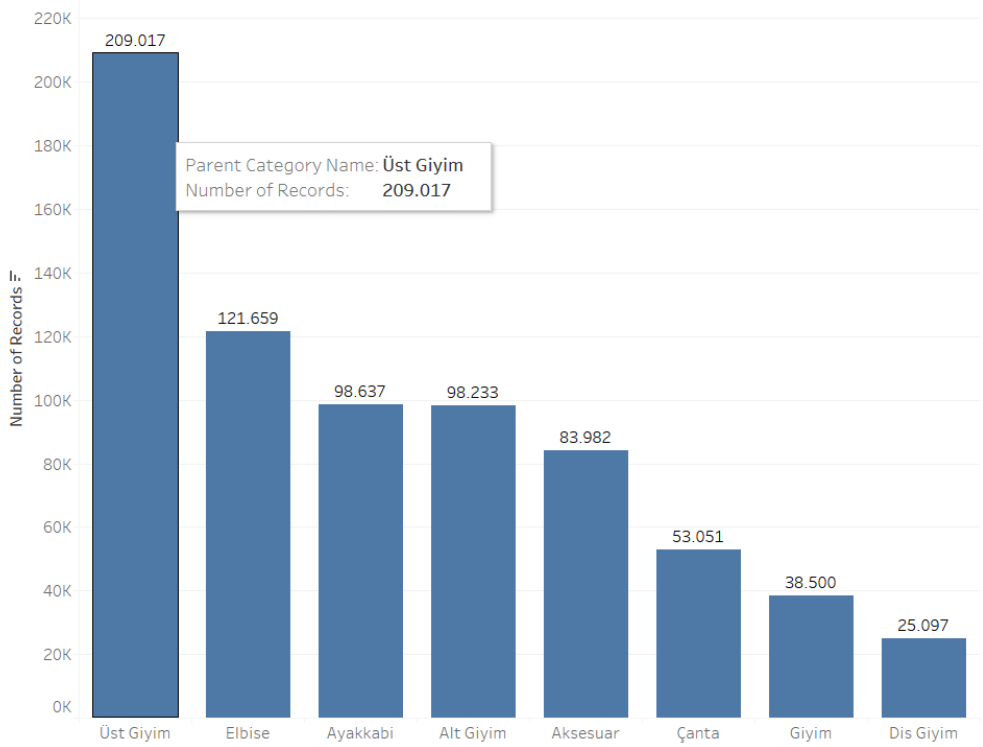



**Figure 7. Product categories**

When looking at the categories, it is seen that the top-wear category has 29% of the products. They are followed by clothing, footwear and bottom categories. In Figure 8, the box plot shows median and outliers in each category.

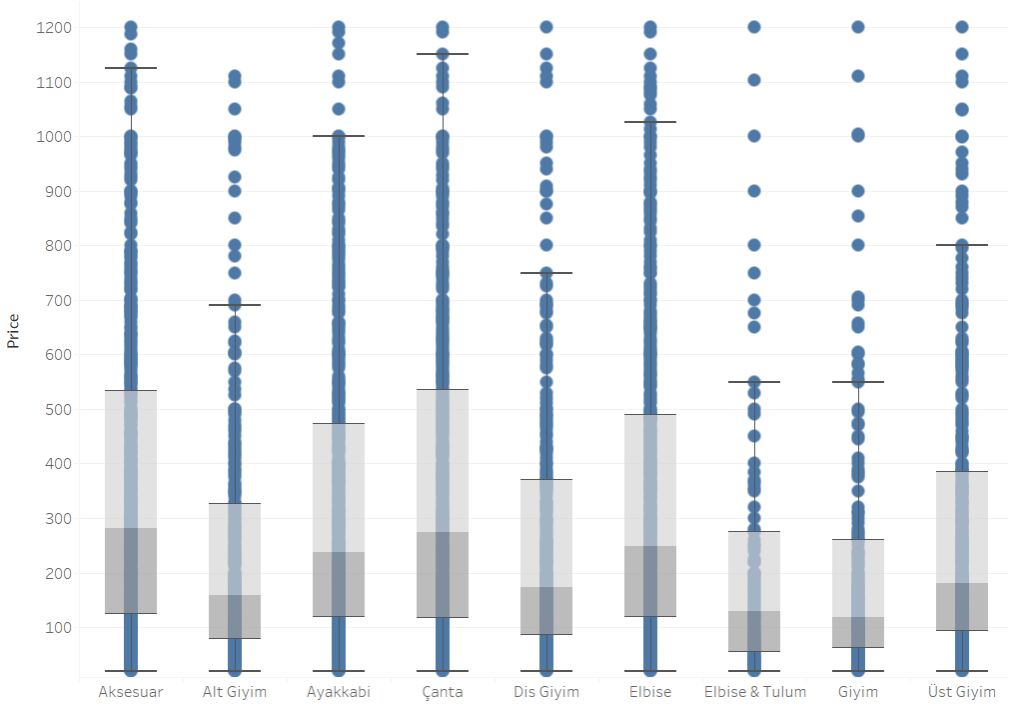



**Figure 8. Price of categories boxplot**

When price was analyzed for each category, the median of accessories category is 283 TL. This category was followed by the bag category with a median of 273 TL. Both two categories are most expensive categories according to sale price.

### 5.1.3. Condition

Condition is using the level of product. Buyer must know how much the product was used according to this information. In addition, product condition affect product sellability. 6.3% of the new tagged products sold in first week. However this rate is 3.4% for the gently worn products. This can show us that the less used products are sold faster. Figure 9 has number of sales by condition in first week of products.

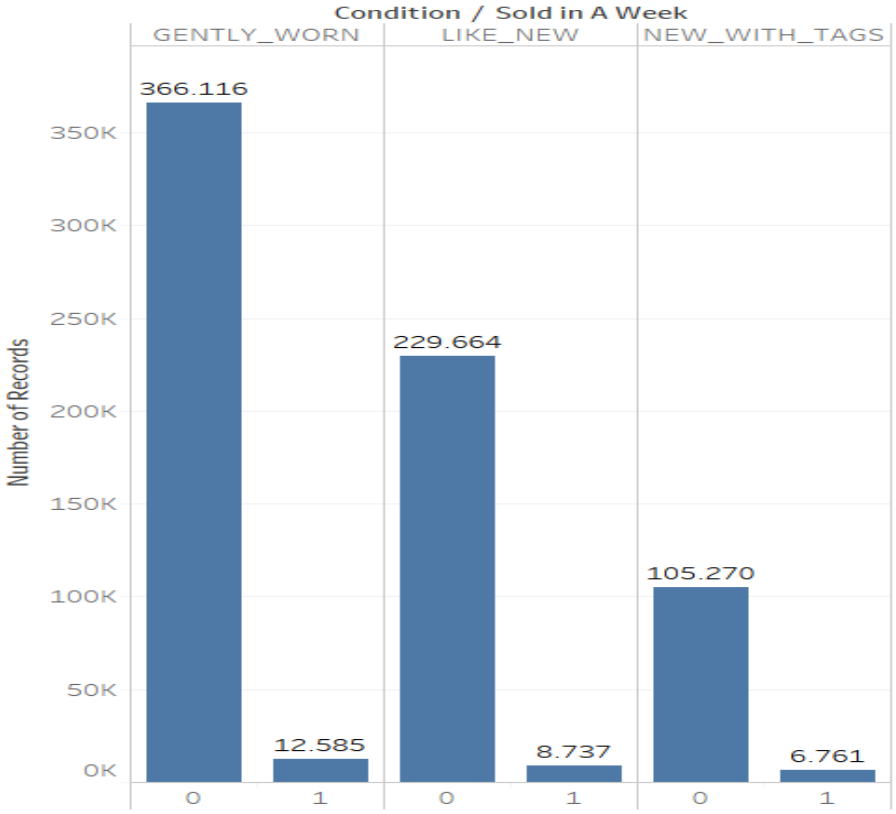



**Figure 9. Number of sold in first week by product condition**

### 5.1.4. Weekly Sales

It is important to know how to obtain the interaction after the products are uploaded to the website. As it is seen in the Figure 10, the number of items sold in the first week is above the number of items sold after the first week. This shows that the buyers demand products in the first week.

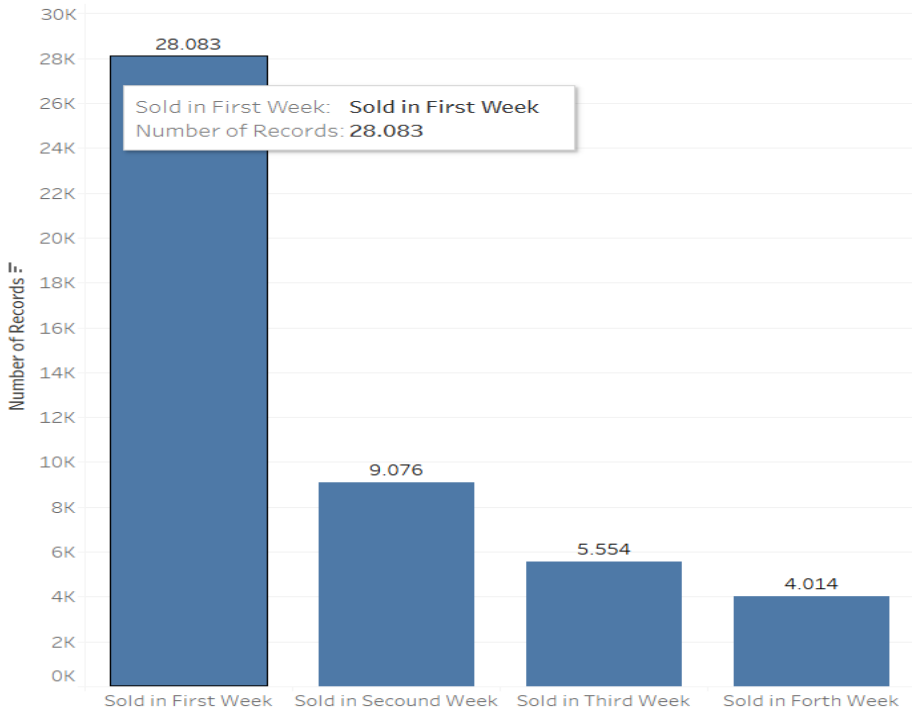



**Figure 10. Weekly sales**

It can also show that products are sold less as their lifetime on the website increases.

## 5.2. Data Cleaning

This section is about handling missing data in the dataset. First, dataset was uploaded to Microsoft Azure ML studio. In a dataset, there are a lot of missing values in "Original Price" columns in the dataset. Because generally users do not want to enter original price of the product on website. Number of missing values in the original price column is 200,322. The distribution of the original price can be seen in Figure 11.


**Figure 11. Frequency of original price**

"Clean Missing Data" object can be used in Azure ML. It can help only deleting rows or replacing the values of some measures. However, it was desired to replace missing values with selling price. Therefore, "Apply SQL Transformation" object was used in Azure in order to replace the N/A values with sales price.

### 5.3. Feature Selection

There are irrelevant features that are not intended to be given to the classification model. For this reason, these properties need to be removed from the data set. The "Select Column in Dataset" object was used in Azure ML Studio. Only the features associated with the model will be moved to the next stage.

The features that are unselected are shown in Figure 12:

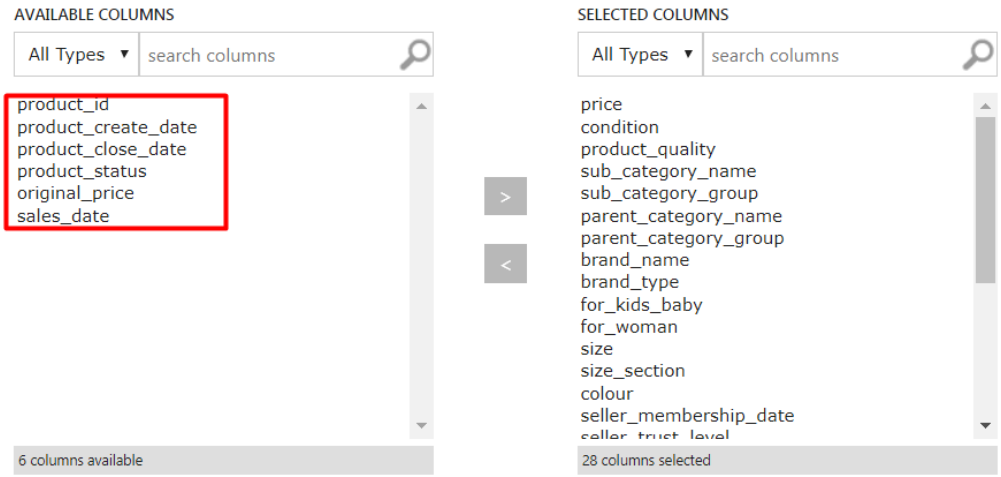



**Figure 12. Unselected features**

There are 27 features in the dataset. It was important to remove attributes that are not related to features. If there are irrelevant features in the model, the model might learn incorrectly. To prevent this, "Filter Based Feature Selection" object was used. Feature Selection refers to the process of applying statistical tests to inputs, given a specified output, to determine which columns are more predictive of the output. "Pearson Correlation" method was used as a feature scoring method and target columns is sales label.

### 5.4. Model Building

In this project, various classification techniques were used. Firstly, the data that has features and labels trained with classification algorithms. After that, new data's labels are used to estimate with this taught model.

In this project, following classification algorithms were tried:

Decision Trees

Logistic Regression

Neural Network

Support Vector Machine

First, the data was need to be split between train and test. “Split Data” object was used in Azure ML. Percentage of the training data was chosen as 70%. The idea is that more training data is preferable because it makes the classification model better whilst more test data makes the error estimate more accurate. Then train data was sent to each algorithm.

As it was known, each model has different initial parameters. But, it was not known which model was the best. For this reason, “Tune Model Hyperparameters” was used for swap parameters. This module performs a parameter sweep over the specified parameter settings, and learns an optimal set of hyperparameters, which might be different for each specific decision tree, dataset, or regression method. The process of finding the optimal configuration is sometimes called tuning.

In this project, most important classification metric is precision. The right marking of sellable products is the main priority. Therefore, the false-negative count should be low while the true-positive count is preferred to be high. For this reason, precision metric was used when training models with “Tune Model Hyperparameters” object. Random sweep method was chosen and maximum number of runs were set as five. If the number is increased, the model runs longer.

Hyperparameters module chooses best parameters for each algorithm. For this reason, the model can be trained with parameters with the highest sensitivity value. Table 4 shows the best parameters for each algorithm.

**Table 4. Algorithm parameters**

| Algorithm | Parameter | Value |
|---|---|---|
| Decision Tree | Number of Leaves | 7 |
| | Minimum Leaf Instances | 2 |
| | Learning Rate | 0.0362 |
| | Number of trees | 37 |
| Logistic Regression | L1Weight | 0.0246 |
| | L2Weight | 0.0129 |
| | Memory Size | 33 |
| Neural Network | Learning Rate | 0.0384 |
| | Loss Function | CrossEntropy |
| | Number of iterations | 27 |
| Support Vector Machine | Number of iterations | 78 |
| | Lambda | 0.000056 |

### 5.5. Model Results

After model building was completed, this section explains results of models. All model results have accuracy score between 0.90 and 0.96. It was known that there are many zero-labeled rows in data set. Before the products are sold, most of them are removed from the website by sellers. This means accuracy score comes from zero-labeled data.

It was needed to pay attention to precision score of each model. Because the main goal is to make accurate prediction for the salable products. Precision is a useful measure of success of prediction when the classes are very imbalanced. Therefore, the models were compared over the precision score. Scores of models are shown in Table 5.

**Table 5. Model scores**

| Model | Accuracy Score | Precision Score | F1 Score | Recall Score |
|---|---|---|---|---|
| Decision Tree | 0.96 | 0.93 | 0.82 | 0.73 |
| Logistic Regression | 0.90 | 0.65 | 0.40 | 0.29 |
| Neural Network | 0.92 | 0.72 | 0.53 | 0.42 |
| Support Vector Machine | 0.90 | 0.63 | 0.37 | 0.26 |

As a result, two-class boosted Decision Tree model has a better precision and accuracy score. It means decision tree can separate more accurately positive labels. Decision tree model can be used for this problem.

Receiver Operating Characteristic curve (or ROC curve.) is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a diagnostic test. It The ROC curve of decision tree model can be seen in Figure 13.

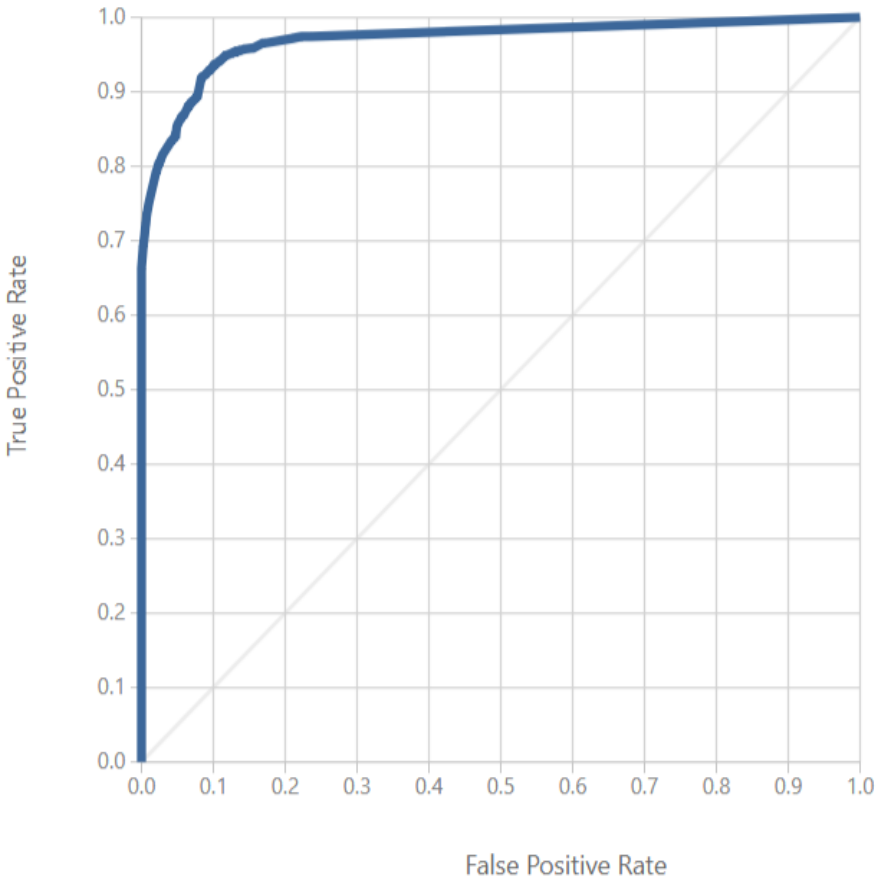



**Figure 13. ROC curve of decision tree**

### 5.6. Feature Importance Analyze

After the appropriate model was selected, the importance of the features in the model will be examined. “Permutation Feature Importance” object was used in Azure ML studio.

In this module, feature values are randomly shuffled, one column at a time, and the performance of the model was measured before and after. It can be chosen as one of the standard metrics provided to measure performance.

The scores that the module returns represent the change in the performance of a trained model, after permutation. Important features are usually more sensitive to the shuffling process, and will thus result in higher importance scores (Martens et. al., 2018).

There are top five features which have highest scores as a result of feature importance analysis. These are shown in Table 6.

**Table 6. Feature importance sore**

| **Feature** | **Score** |
|---|---|
| Number of Likes | 0.0288 |
| Number Of Bids | 0.0115 |
| Number Of Comments | 0.0071 |
| Seller’s Feedback Avg. | 0.0017 |
| Number Of My Brand User | 0.0007 |

# 6. CONCLUSION

In this study, the main goal is to determine the features that affect the sellability of the products on an e-commerce website and to boost the sales using these features. This was done by using self-learning intelligent models. Products were shown to customers on the main page of the website according to certain assumptions in current structure. This can be revised by using self-learning intelligent models.

First of all, the data need to be understood to solve the problem. Some features in the dataset were examined with descriptive statistical methods such as histograms, scatter plots and boxplots. This phase was about discovering the data.

As it was seen during the data discovery, the number of products from some categories were almost double of the other products. Also, gently used products have much higher sellability than the others. It was found that 30% of total sold products were sold in the first week. This shows that the product gets the most interaction when it is first uploaded to the site. After these inferences, the data was understood better and the sellability of products prediction model was ready to be built.

Four different supervised classification models were used to estimate the sellability of the products. Among these models, “Decision Tree Algorithm” was chosen because it has the highest accuracy and precision score as explained in the results section of this study. At the end of the study, it was found the most accurate model of the sellability of the products is the “Decision Tree Algorithm”.

After the model has been put into the production environment, the most important and influential features needed to be found in order for the model to perform better. In the feature importance analysis, it was seen that the calculated variables like number of bids and number of likes are more critical than other features. This shows how important feature engineering is. At the same time, interaction with products such as likes, comments or bids are more meaningful than the product base features like brand and color. For this reason, marketing teams should analyze and follow the movements of the products more closely.

After this study, next step is determining which one of the salable products should be displayed on the main page of the website. Because there is limited space on main page of website and in ad spaces. To determine the product to be displayed, the model should be developed and sellability scores should be calculated as future work.

# 7. APPENDIX

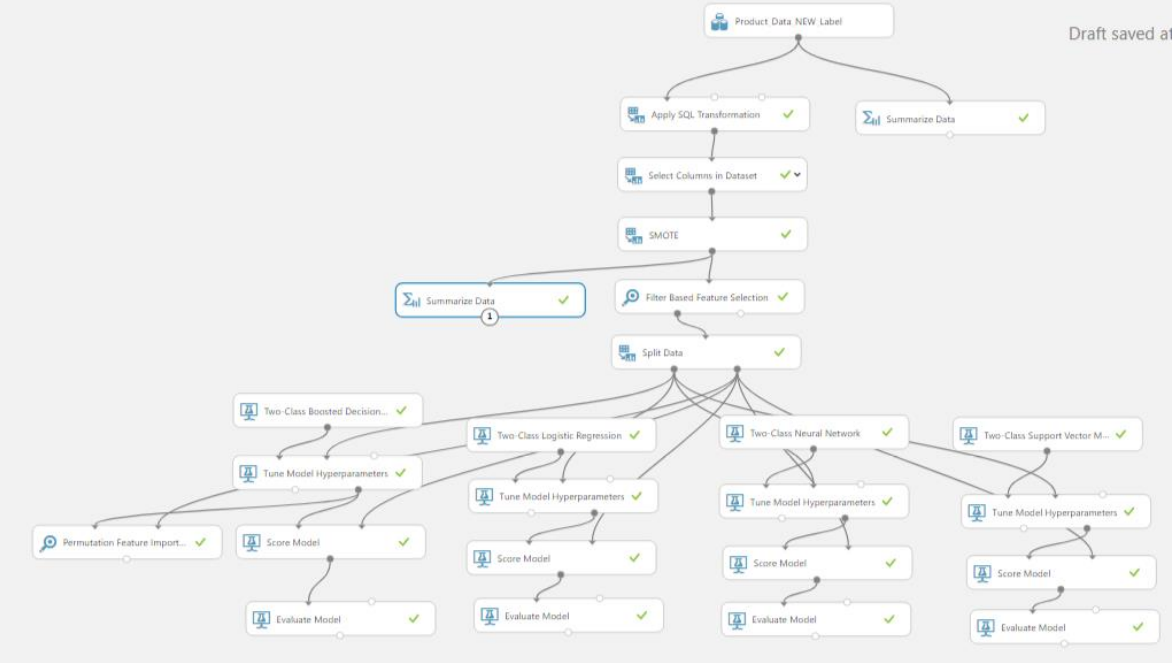
## SQL Script (Structured Query Language) for data preparation

```
SELECT
p.id    AS PRODUCT_ID,
date ( p.created_date)      AS PRODUCT_CREATE_DATE,
date ( CASEWHENproduct_status=    'CLOSED'  THENupdated_date      ELSE NULL end)   AS
PRODUCT_CLOSE_DATE,
p.product_status,
p.price,
p.original_price,
p. "condition"     ,
p.product_quality,
coalesce  (sc.title,       'NA' )   AS SUB_CATEGORY_NAME,
coalesce  (sc.category_group,      'NA' )   AS SUB_CATEGORY_GROUP,
coalesce  (pc.title,       'NA' )   AS PARENT_CATEGORY_NAME,
coalesce  (pc.category_group,      'NA' )   AS PARENT_CATEGORY_GROUP,
coalesce  (b.title,       'NA' )   AS BRAND_NAME,
coalesce  (b.BRAND_TYPE,'NA' )   AS BRAND_TYPE,
b.for_kids_baby      AS for_kids_baby,
b.for_woman    AS for_woman,
coalesce  (s.title,       'NA' )   AS SIZE,
coalesce  (s.SIZE_SECTION, 'NA' )   AS SIZE_SECTION,
coalesce  (col.title,        'NA' )   as   COLOUR,
date (mo.membership_date)     AS SELLER_MEMBERSHIP_DATE,
coalesce  ( mo.seller_trust_level,         'NA' )   AS SELLER_TRUST_LEVEL,
coalesce  (mo.feedback_avg,    0)   AS SELLER_FEEDBACK_AVERAGE,
coalesce  (pr.NOF_REPORTER,0)   AS NOF_REPORTER,
coalesce  (bcap.average_price,      0)   AS AVERAGE_PRICE_OF_BRAND,
coalesce  (co.NOF_COMMENT,0)   AS NOF_COMMENT,
coalesce  (fo.NOF_FOLLOWER_FOR_SELLER,0)AS NOF_FOLLOWER_FOR_SELLER,
coalesce  (li.NOF_LIKES,    0)   AS NOF_LIKES,
coalesce  (block.SELLER_NOF_BLOKED_BY_OTHERS,0)AS SELLER_NOF_BLOKED_BY_OTHERS,
coalesce  (my_b.NOF_MY_BRAND_BY_USER,0)AS NOF_MY_BRAND_BY_USER,
coalesce  ( my_s.NOF_MY_SIZE_BY_USER,0), AS NOF_MY_SIZE_BY_USER,
coalesce  (bid.NOF_BID,  0)   AS NOF_BID,
CASEWHENpoi.product_id       IS   NOTNULL AND DATE_PART('day'   , SALES_DATE -
p.created_date)<    8   THEN1   ELSE 0   ENDAS SALES_LABEL,
CASEWHENpoi.product_id       IS   NOTNULL THENSALES_DATEELSE null     ENDAS
SALES_DATE
FROM
public.product p
INNER JOIN  public.    member mo ON mo.id=p.owner_id
LEFT OUTERJOIN  public.brand b       ON p.brand_id=b.id
LEFT OUTERJOIN  public.    size    s   ON p.size_id=s.id
LEFT OUTERJOIN  public.category sc         ON p.category_id=sc.id
LEFT OUTERJOIN  public.category pc         ON sc.parent_id=pc.id
LEFT OUTERJOIN  public.product_colour pcol           ON p.id=pcol.product_id
LEFT OUTERJOIN  public.colour       AS col    ON pcol.colour_id=col.id
LEFT OUTERJOIN  public.brand_category_avg_price          bcap   ON
p.brand_id=bcap.brand_id       ANDp.category_id=bcap.category_id          AND
p. "condition"     =bcap.product_condition
LEFT OUTERJOIN  (
SELECTproduct_id,    max(created_date)      AS SALES_DATE
```

```
FROM public.product_order_item
WHERE order_item_status NOT IN ('FAILURE', 'CANCELLED', 'WAIT_PAYMENT')
AND order_item_fraud_status= 'NOT_FRAUD'
GROUP BY product_id
) AS poi ON poi.product_id=p.id
LEFT OUTER JOIN (
SELECT reported_product_id, count(DISTINCT reporter_id) AS NOF_REPORTER
FROM public.product_report
GROUP BY reported_product_id
) pr ON p.id=pr.reported_product_id
LEFT OUTER JOIN
(
SELECT product_id, count(*) AS NOF_COMMENT
FROM public."comment"
WHERE deleted= 'false'
GROUP BY product_id
) AS co ON p.id=co.product_id
LEFT OUTER JOIN
(
SELECT followee_id, count(DISTINCT follower_id) AS NOF_FOLLOWER_FOR_SELLER
FROM public.follow_member
WHERE deleted= 'false'
GROUP BY followee_id
) AS fo ON mo.id=fo.followee_id
LEFT OUTER JOIN
(
SELECT product_id, count(DISTINCT liker_id) AS NOF_LIKES
FROM public.likes
WHERE deleted= 'false'
GROUP BY product_id
) AS li ON p.id=li.product_id
LEFT OUTER JOIN
(
SELECT blocked_id, count(DISTINCT blocker_id) AS SELLER_NOF_BLOKED_BY_OTHERS
FROM public.member_block
WHERE deleted= 'false'
GROUP BY blocked_id
) block ON mo.id=blocked_id
LEFT OUTER JOIN
(
SELECT brand_id, count(DISTINCT member_id) AS NOF_MY_BRAND_BY_USER
FROM public.my_brand
WHERE modifier= 'MEMBER'
GROUP BY brand_id
) AS my_b ON p.brand_id=my_b.brand_id
LEFT OUTER JOIN
(
SELECT size_id, count(DISTINCT member_id) AS NOF_MY_SIZE_BY_USER
FROM public.my_size
WHERE modifier= 'MEMBER'
GROUP BY size_id
) AS my_s ON p.size_id=my_s.size_id
LEFT OUTER JOIN
(
SELECT product_id, count(*) AS NOF_BID
FROM public.product_bid
GROUP BY product_id
```

```
) AS bid ON p.id=bid.product_id
WHERE 1=1
AND product_status NOT IN ('DISAPPROVED','WAITING_APPROVAL')
AND DATE(p.created_date)>= '2018-06-30' AND DATE(p.created_date)< '2018-07-01'
```

## Microsoft Azure ML Studio Pipeline



# REFERENCES


Anyanwu, N. and Shiva, G. (2009). Comparative Analysis of Serial Decision Tree Classification Algorithms. International Journal of Computer Science and Security. 3, pp. 230-240

Chen, S. (2014). E-Commerce Sales Prediction Using Listing Keyword.

Gelman, A. and Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

Joshi, R. (2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. Retrieved from http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Kumar, A. (2018). Introduction to Machine Learning. Retrieved from http://www.allprogrammingtutorials.com/tutorials/introduction-to-machine-learning.php

Liu, X. and Li, J. (2016). Using Support Vector Machine for Online Purchase Predication.

Lui, H. (2018). What Is the Future of E-commerce in 2018 and Beyond? 10 Trends. Retrieved from https://www.shopify.com/enterprise/the-future-of-ecommerce

Martens, J., Petersen, T., Astala, R. et. al. (2018). Permutation Feature Importance. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance

Tekel, S. (2014). E-commerce Organizations and Turkey. European Journal of Research on Education. 2 , pp 25-33

Steinweg-Woods, J. (2016). A Guide to Gradient Boosted Trees with XGBoost in Python. Retrieved from https://jessesw.com/XG-Boost/