Capstone Project: Sales Prediction of Summer Clothes in Wish.com

Akhil Malkani, Rutgers Business School, RUID- 190003125, Email: [akhil.malkani@rutgers.edu](mailto:akhil.malkani@rutgers.edu)

A picture containing text, different, several

Description automatically generated

1. **INTRODUCTION**

Studying top products requires more than just product listings. You also need to know what sells well and what does not. The online marketing space is in constant shift as new technologies, services, and marketing tactics gain popularity and become the new standard. Online store owners are one of the many different segments affected by these constant evolutions. In order for these business owners to survive and thrive, they need to be able to make better decisions faster. This is where Data Analytics comes into play.

In this Data analytics process we will be gathering data from all areas that have an impact on the online store and use this information to understand the trends and the shift in consumers’ behavior to make data-driven decisions that will drive more online sales.

The evolution in consumer demand, combined with technological innovations, will continue to drive growth in global ecommerce sales. According to Statista, the [number of people buying goods and services online](https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/) is expected to reach 2.14 billion in 2021, up from 1.66 billion global digital buyers in 2016.

What’s even more impressive is the fact that the industry is forecasted to double in size within the next two years and grow from 3.53 trillion US dollars in retail ecommerce sales in 2019, up to [6.54 trillion US dollars in 2022](https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/).

The key drivers of success over the next decade will be centered on building a deep understanding of and connection to the empowered consumer, and the only way to understand consumer behavior is to measure and analyze.

Besides Analyzing data We are also going to build a model that can help predict how well a product is going to sell, i.e., the exact sales for each product.

Such a model has many implications and could be used in many different ways, the most straightforward being to adjust how much of a product should be kept in stock.

1. **DATA SOURCE & DESCRIPTION**

The dataset was downloaded from Kaggle.com (<https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish>) . The data comes from the [Wish](https://www.kaggle.com/jmmvutu/wish.com) platform.  
Basically, the products listed in the dataset are those that would appear if you type "summer" in the search field of the platform.

Info on Dataset:

1. title : Title for localized for european countries. May be the same as title\_orig if the seller did not offer a translation.
2. title\_orig : Original english title of the product.
3. price : price for the buyer
4. retail\_price : Retail price, or reference price in other stores/places. Used by the seller to indicate a regular value or the price before discount.
5. currency\_buyer : currency of the prices
6. units\_sold : Number of units sold. Lower bound approximation by steps
7. uses\_ad\_boosts : Whether the seller paid to boost his product within the platform (highlighting, better placement or whatever).
8. rating : Mean product rating.
9. rating\_count : Total number of ratings of the product
10. rating\_five\_count : Number of 5-star ratings (there are also similar rating columns for four, three .. stars)
11. badges\_count : Number of badges the product or the seller have.
12. badge\_local\_product : A badge that denotes the product is a local product. Conditions may vary (being produced locally, or something else). Some people may prefer buying local products rather than. 1 means Yes, has the badge.
13. badge\_product\_quality : Badge awarded when many buyers consistently gave good evaluations 1 means Yes, has the badge
14. badge\_fast\_shipping : Badge awarded when this product's order is consistently shipped rapidly
15. tags : tags set by the seller
16. product\_color : Product's main color
17. product\_variation\_size\_id : One of the available size variation for this product
18. product\_variation\_inventory : Inventory the seller has. Max allowed quantity is 50
19. shipping\_option\_price : shipping price
20. shipping\_is\_express : whether the shipping is express or not. 1 for True
21. countries\_shipped\_to : Number of countries this product is shipped to. Sellers may choose to limit where they ship a product to
22. inventory\_total : Total inventory for all the product's variations (size/color variations for instance)
23. has\_urgency\_banner : whether there was an urgency banner with an urgency
24. merchant\_rating : merchant's rating

Graphical user interface, application, table

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface, application, Teams

Description automatically generated

1. **Problem Statement**

Topics covered and questions to answer from the data:

* Comparison between price and retail price
* Sales versus origin country
* Sales comparison by colors
* Sales comparison by ratings of products
* What factor contributes most to a fast shipping badge?
* Tags encoding
* Which badge contributes most to the sales of a product?
* What kind of merchants are likely to gain product success?
* The main task is to predict sales of a product.

1. **Methodologies**
2. **Data Pre-Processing**

Data pre-processing is the process of converting raw data into a well-readable format to be used by a machine learning model. This method includes three important steps which are data cleaning, data transformation and feature selection.

**Data cleaning** - is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set’s accuracy without necessarily deleting information.

**Data Transformation** - is the process of changing data from one particular format or arrangement to another one. The data transformation tools and methods are important since information can reside in different locations and has a plethora of formats, and organizations must have access to solutions for converting this diverse information depending on the unique needs of its business ecosystem. The end goal of this process is to make data more readable when it is moved from one application or database to another.

**Feature Selection** - is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

Data Preprocessing is a key step in the path to making models that can predict/classify depending on the dataset we have and the question we aim to answer. At some level, this requires you to be aware of the background of your data and the question you intend to answer. There is a lot of underlying inferences that can be extracted while answering the actual question. This is possible due to the data preprocessing phase.

I have looked up quite a few online articles about what different methods are involved and when should we apply what. It has been an interesting read and I have attempted to apply what I learnt here.

Below are some steps that are used at times in the Data Preprocessing stage. The list is not exhaustive, these are some of the methods that were used in this dataset

1. [Removing the null values](https://www.kaggle.com/mani97/sales-prediction-preprocess-boosting-77#1)
2. [Transform categorical variables](https://www.kaggle.com/mani97/sales-prediction-preprocess-boosting-77#2) (as necessary)
3. [Removing the features that have 1 unique value](https://www.kaggle.com/mani97/sales-prediction-preprocess-boosting-77#3)
4. [Engineer new feature](https://www.kaggle.com/mani97/sales-prediction-preprocess-boosting-77#4) (if there is a possibility)
5. [Remove unnecessary features](https://www.kaggle.com/mani97/sales-prediction-preprocess-boosting-77#5)
6. Binning data (if required) (was not required here)
7. **Exploratory Data Analysis**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them

The purpose of exploratory data analysis is to:

1) Check for missing data and other mistakes.

2) Gain maximum insight into the data set and its underlying structure.

3) Uncover a [parsimonious model](https://www.statisticshowto.com/parsimonious-model/), one which explains the data with a minimum number of [predictor variables](https://www.statisticshowto.com/independent-variable-definition/).

4) Check assumptions associated with any model fitting or [hypothesis test](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/).

5) Create a list of [outliers](https://www.statisticshowto.com/find-outliers/)or other anomalies.

6) Find [parameter](https://www.statisticshowto.com/what-is-a-parameter-statisticshowto/)estimates and their associated [confidence intervals](https://www.statisticshowto.com/probability-and-statistics/confidence-interval/)or [margins of error](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/margin-of-error/).

Identify the most influential [variables](https://www.statisticshowto.com/variable/).

1. **Data Modelling and Prediction**

Many times when people think of modeling data, they directly think of regression models to generate predictions of some target variable based on gathered data. However, classification models are extremely useful for generating predictions based on data as well, but in a different way. Regression models are used to predict continuously distributed target variables, which have an uncountably infinite number of outcomes. Examples of regression problems include predicting the price of a house at sale, predicting the temperature on a given day, and predicting how many points an NBA player will score in their next game. In classification data models, the target variable we are trying to predict has a discrete distribution, which has a finite number of outcomes. Examples of classification problems include predicting which candidate will win an election and predicting the sales like we are going to do in this project.

A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

We need to divide the dataset into training set and test set. The training set is be preprocessed, and each model is trained and validated using cross-validation. During this process, we put the test set aside and don't even look at it to make sure the model is unbiased. Once the model type and hyperparameters have been selected, the generalized error is measured on the test set.

The Classification Models used in the project are:

1. **Gradient Boosting**- is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.
2. **Decision tree**- builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
3. **Random forest**- is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
4. **AdaBoost-** is best used to boost the performance of decision trees on binary classification problems. It can be used to boost the performance of any machine learning algorithm. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.
5. **SVC**- is part of Support Vector Machine(SVM) Model which is a supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The idea of SVM is simple, the algorithm creates a line or a hyperplane which separates the data into classes.
6. **K-Nearest Neighbor**- also known as KNN is a supervised learning algorithm that can be used for regression as well as classification problems KNN works on a principle assuming every data point falling near to each other is falling in the same class.
7. **Experiments**

I started the Pre-Processing Method with cleaning the dataset, where I want to get rid of all duplicate and treat null values.

Table

Description automatically generated

* The rating count has zero null values, however individual rating count has 45 null values. Does this mean that rating count of 45 products are zero?
* Product color and size variation has 41 and 14 null values respectively. Perhaps this products have no variations?
* has\_urgency banner and urgency\_text have similar null values. This makes sense, however, I will convert this to zeros and ones.
* Origin countries have 17 null values. I might replace this with the mode if necessary.
* Surprisingly there are 4 missing merchant names but only one missing info subtitle.
* There are 1347 missing merchant profile pictures. I am expecting this to match with 'zeros' in 'merchant\_has\_profile\_picture'

**Product\_Color** has 41 null values and has 101 colors, some of them are redundant. One option to deal with this would be to aggregate all the colors into similar groups (for example denimblue goes into blue), to reduce the number of categories.

Table

Description automatically generated

Replaced the Nan values with others as they could be a shade that’s not in their color palette or for some reason not mentioned.

Aggregating All the colors into a similar group helped with data analysis and gave a clear picture on which are popular color choices as one can see from the analysis black and white are the most popular choices.

A picture containing table

Description automatically generated

Table

Description automatically generated

Chart, bar chart

Description automatically generated

This shows although white and black are produced more, the colors that are performing well are more vibrant.

**Origin\_Country** has 16 null values, and majority of products are from China and US.

Chart, bar chart

Description automatically generated

Since other countries have less contribution for the set, those items along with nan were replaced with 'Other'.

**product\_variation\_size\_id** has 14 null values and lot of these sizes are of different scale so I converted them to the same scale.

Text, letter

Description automatically generatedText

Description automatically generated

Table

Description automatically generated

Chart, bar chart

Description automatically generated

Majority of Sales are of small size so it makes sense for the company to increase production of small size clothes to meet the demand.

**Rating\_count**

rating\_one\_count rating\_two\_count rating\_third\_count rating\_fourth\_count rating\_five\_count rating rating\_count The first 5 columns have missing values. They turned out to be products without any rating, although for some reason the rating of these products is 5, even though they have a rating\_count of 0.

For these products, I will change the rating and all the rating counts to 0, indicating that they did not get any votes yet.

Text

Description automatically generated

**Has\_urgency\_banner** has 1071 null values. The column has\_urgency\_banner has values either 1 or nan, so I will turn the nan values of this column into 0.

A picture containing text

Description automatically generated

merchant\_profile\_picture contains the url to merchants profile picture where more than 80% of data is missing

It seems that there are only urgency\_text in French, one indicating limited quantity and another indicating price reduction. Only one of them is "Réduction sur les achats en gros" most of it is in French and not useful to us. Hence we will discard both.

After Dealing with all the null values, I created new columns for better understanding the relation between retail price and price.

Graphical user interface, text, application

Description automatically generated

We will check out the tags to get more information of what is popular on wish platform.

Chart, histogram

Description automatically generated

As we can see womens fashion ,summer, fashion and majority of popular tags are related to female gender.

To get more insight I created a column tag\_quality which indicates how many popular tags contain on a specific product tag

Graphical user interface, text, application, email

Description automatically generated

I created Tags\_Count Column to check the number of tags included in the column.

Graphical user interface, text, application, email

Description automatically generated

Visually you can see that most of the time a product has around 0-20 tags on them.

Shape

Description automatically generated

I created another column title\_localised to check whether the title was transferred or if its orginial . 0 is that its original and localized whereas 1 is transferred. You can see most of the title’s are transferred or they change the title name.

Chart, bar chart

Description automatically generated

I carried on with my experiment to find relationship between price and sale. Retail price is used by the seller to indicate a regular value or the price before discount. How do price, retail price and the price drop in discount define the product success?

From my analysis I found:

1.There are some cases where the price is low still the units sold are below average, possible reasons the product might not be good enough as per the buyers or there are some other factors affecting the price we haven't touched yet

There is a visible downward trend in units sold as the price increases. Products with high sales are usually concentrated in the price range of 0-20.

The difference between actual price and retail price is quite large. Prices are more concentrated while the retail prices have more outliers. This could be a popular sales strategy.

The steep price drops don't necessarily result in product success.

Chart, scatter chart

Description automatically generated

2.Median of units sold is 1000, by this we can consider that products with units sold below 1000 (inclusive) were below average and products with units sold are very successful. # It totally depends on your business goals which price range you want to focus on.

Median for rating is 3.85 and the products in top selling cluster has rating between 3.35 to 4.1 seems very reasonable Rating is very important to determine the potential of product Still there are some products with 5 star rating yet unable to cross the 100-1000 unit sold line there are some really bad performing products with rating below 3

Chart, box and whisker chart

Description automatically generated

I created a correlation heatmap to see which columns has what effect on unit\_sold to choose the right columns for our model, since we don’t want information which has no effect on unit\_sold,

Chart

Description automatically generated

Major Factors which increases the 'units\_sold'

* Rating Count
* Rating
* Rating Five Count
* Rating Four Count
* Rating Three Count
* Merchant Rating Count
* Merchant Rating
* Merchant Has Profile Picture
* Is Title Localised
* Rating
* Badge Product Quality
* Badges Count
* Product Variation Inventory
* Number of Tags
* Title Localised
* Number Of Tags
* Have Rating on Merchant Subtitle
* High Retail Price
* High Buyer Price

Major Factors which decreases the 'units\_sold'

* Rating Two Count
* Rating One Count
* Price Increment
* Shipping Price Increment
* If Merchant title and Merchant name has differnt names
* No Urgency Banner
* High Price (Buyer Price)
* Lower discount rate (Buyer price compare to retail price)

After going through the heatmap the next step is to remove unwanted columns.

rating\_count parameter is a driving factor for the unit\_sold. It's obvious that rating\_count increases the model accuracy. Since the model is predicting future sales we do not know rating\_count for the model.

**merchant could buy enough stock to maximize the revenue or stop buying way too much of a product. So the below model could be used, If the merchant has some data about the first few weeks of sales**.

Columns: title, title\_orig, merchant\_profile\_picture, product\_url, product\_picture, product\_id, merchant\_id, merchant\_info\_subtitle, merchant\_name, merchant\_title, shipping\_option\_name, urgency\_text

These will be dropped for now, as the likelihood of these affecting the number of units sold is pretty less. For some of the features present above, a corresponding feature already exists in the dataset that provides more information relevant to the model we want to make.

The rating count column will also be removed for now as we already have features of the distribution of rating count across (5/4/3/2/1) which gives us a more detailed information than 'rating count'

We will use the Scikit-learn to create a prediction models and choose the model which has the highest accuracy.

Below are the results of all the models I created using Scikit-learn.

Table

Description automatically generated

1. **Related Work**

My model is inspired by what I was taught in my Data Mining class at Rutgers MITA program.

I referred to the Data Mining lecture slides provided by Prof. Meng Qu and used various sites such as:

<https://towardsdatascience.com/>

<https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>

<https://scikit-learn.org/stable/index.html>

<https://www.udemy.com/course/complete-python-bootcamp/>

[www.kaggle.com](http://www.kaggle.com)

1. **Conclusions**

* Majority of the products are black and white. This might have been defined wrong by the merchants. If that’s not the case, the merchants can be encouraged to include more variation to these to increase its sales since the sale are more of vibrant colors in summer than white and black
* The site mainly sells female clothing as we have seen from the popular tags like womens clothing and women are most used on products.
* Higher units sold means higher rating count, hence it is advisable to suggest brands to encourage people to give rating and provide nice customer service to earn higher rating.
* The use of ad boosts does not seen to have any effect on the units sold and the site may lose revenue from this ads.
* More detailed units sold and inventory levels would have been more helpful for analysis.
* Product quality badges seems to increase the success of the products. Perhaps reviewing more products will increase motivation for merchants to improve their product? Different levels of badges can be applied?
* The tags can be improved so that products can be categorized more specifically. This can be done by reducing the number of tags per product, so the merchants are forced to choose their tags more wisely.
* Gradient Boosting is the model chosen to predict sales, I believe this model could help ecommerce platforms predict sales and make critical business decision based on prediction.