REPORT – CAPSTONE PROJECT II

Muzaffer Estelik
estelik.muzaffer@gmail.com

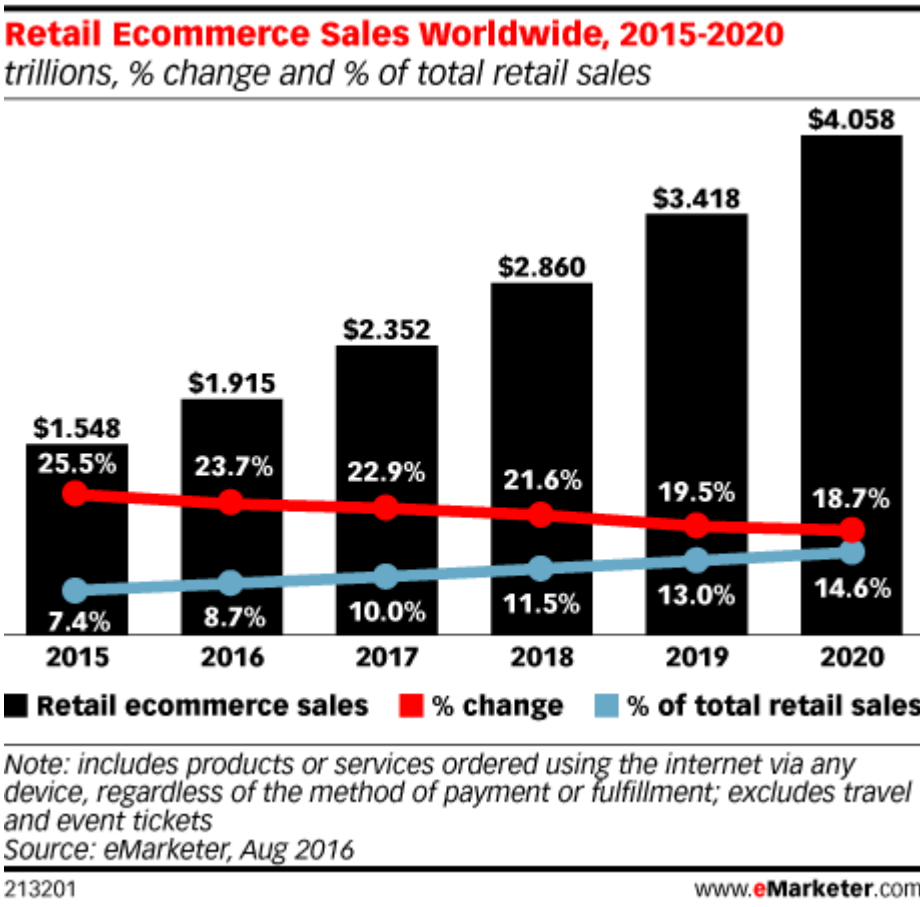
1. PROBLEM

Prediction of ratings based on women's e-commerce reviews.

a. Introduction:

E-**commerce** is the activity of buying or selling of products on online services or over the Internet[1]. Ecommerce refers to commercial transactions conducted online. This means that whenever you buy and sell something using the Internet, you're involved in e-commerce.

Needless to say, e-commerce has grown by leaps and bounds since the first time e-commerce transaction was committed in 1994. BigCommerce cites that e-commerce is growing 23% year-over-year, and according to eMarkerter, global e-commerce sales are expected to top $27 trillion in 2020 — and that's just statistics for the retail sector.



1 https://en.wikipedia.org/wiki/E-commerce

REPORT – CAPSTONE PROJECT II

Global retail e-commerce sales are projected to reach $27 trillion by 2020. It's obvious that, each day e-commerce will be more and more important for the companies.

This project aims using NLP techniques, different machine learning models and Deep Learning for predicting the rates of the products into 2 categories (good, bad).

2. DATA WRANGLING

a. Examining the Features and Samples

Features:

F.Nu	Name	Explanation	Type
1	Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed	Linear
2	Age	Positive Integer variable of the reviewers age	Linear
3	Title	String variable for the title of the review	Text
4	Review Text	String variable for the review body	Text
5	Rating	Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best	Linear
6	Recommended IND	Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended	Binary
7	Positive Feedback Count	Positive Integer documenting the number of other customers who found this review positive	Linear
8	Division Name	Categorical name of the product high level division	Text
9	Department Name	Categorical name of the product department name	Text
10	Class Name	Categorical name of the product class name.	Text

b. Class Distribution:

The ratings are scored by the customer from 1 Worst, to 5 Best. I added up a new feature (Positively Rated) splitting these ratings to a binary classification adding rating scores 1,2,3 as bad (0) and 4,5 as good (1).

REPORT – CAPSTONE PROJECT II

There are 23486 rows, each representing a review for a different clothing item.

There are 10 attributes like clothing id, age, title, review text, etc. The data set is labeled with 2 different classes. Class 1 corresponds to positively rated item. Class 0 corresponds to negatively rated item.

The data set is looking like biased towards positive ratings but there are enough samples of negative ratings for ML models to make predictions.