

**Ratings Prediction Project**

Submitted by:

N THARANI

**ACKNOWLEDGMENT**

Various research papers and websites are used. Research papers like [An Automated Toxicity Classification on Social Media](https://link.springer.com/article/10.1057/s41264-020-00085-7#auth-K__Ramanathan-Kalimuthu) paper by [Ahmad Alsharef](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alsharef%20A%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Karan Aggarwal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Aggarwal%20K%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Sonia](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sonia%20%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Deepika Koundal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Koundal%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168), [Hashem Alyami](https://www.ncbi.nlm.nih.gov/pubmed/?term=Alyami%20H%5BAuthor%5D&cauthor=true&cauthor_uid=35211168) and [Darine Ameyed](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ameyed%20D%5BAuthor%5D&cauthor=true&cauthor_uid=35211168); A New Application of Social Impact in Social Media for Overcoming Fake News paper by [Cristina M. Pulido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Pulido%20CM%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Laura Ruiz-Eugenio](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ruiz-Eugenio%20L%5BAuthor%5D&cauthor=true&cauthor_uid=32260048), [Gisela Redondo-Sama](https://www.ncbi.nlm.nih.gov/pubmed/?term=Redondo-Sama%20G%5BAuthor%5D&cauthor=true&cauthor_uid=32260048) and [Beatriz Villarejo-Carballido](https://www.ncbi.nlm.nih.gov/pubmed/?term=Villarejo-Carballido%20B%5BAuthor%5D&cauthor=true&cauthor_uid=32260048); How Malicious Comments on Social Media Has a Huge Impact on the Negative Attitude paper by Puiyu Huang; How Vicious Do You Think Your Social Media Comments Are? article by [Navjot Bians](https://medium.com/@biansnavjot?source=post_page-----204128141eab--------------------------------); Why People Post Benevolent and Malicious Comments Online paper by So-Hyun Lee, Hee-Woong Kim, Review rating prediction using combined latent topics and associated sentiments: an empirical review by [Anbazhagan Mahadevan](https://link.springer.com/article/10.1007/s11761-019-00278-6#auth-Anbazhagan-Mahadevan), [Michael Arock](https://link.springer.com/article/10.1007/s11761-019-00278-6#auth-Michael-Arock), etc., are used. Website like researchgate.net, geeksforgeeks, etc., are used as references. The data is received from the client which is their own data.

Thanking SWATI MAHASETH, my guide from FLIPROBO TECHNOLOGIES for clearing all my doubts while undergoing the project.

**INTRODUCTION**

* Business Problem

Product attributes, average consumer ratings, and single affect-rich positive or negative consumer reviews influenced hypothetical online purchasing decisions of younger and older adults. The preference for the higher-rated product, however, could be overridden by a single affect-rich negative or positive review. These issues suggest that people do not consider aggregated consumer information and positive reviews focusing on positive experiences with the product, but are easily swayed by reviews reporting negative experiences.

In order to understand the effect of online reviews on consumer purchase behavior, more than 20000 online reviews are collected. We want to predict ratings for the reviews which were written in the past and they don’t have a rating. Our goal is to build a prototype to build an application which can predict the rating by seeing the review.

* Conceptual Background of the Domain Problem

Systematic reviews and meta-analyses have become increasingly important. As with all research, the value of a systematic review depends on what was done, what was found, and the clarity of reporting. Several early studies evaluated the quality of review reports. Several approaches have been developed to conduct systematic reviews on a broader array of questions. This paper studies the influence of ratings according to online reviews of experience goods from a new perspective of consumer learning.

The moderate reviews, negative reviews, logistics rating and service rating are not significant in the results. Only ratings based on the reviews is taken into consideration. As consumers’ guide information, anonymous transmission of online reviews make more consumers are willing to provide their own real-life experience, even if it is a negative one. Review rating of goods is the initial attitude of consumer, which is evaluated by the consumer on goods which is taken into point and the model is made.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Certain websites and papers that helped me to take insights from are:

1. Influence of consumer reviews on online purchasing decisions by [Bettinavon Helversena](https://www.sciencedirect.com/science/article/pii/S0167923618300861" \l "!), [KatarzynaAbramczukb](https://www.sciencedirect.com/science/article/pii/S0167923618300861#!)
2. An emerging consensus on rating quality of evidence and strength of recommendations by [Gordon H Guyatt](https://www.ncbi.nlm.nih.gov/pubmed/?term=Guyatt%20GH%5BAuthor%5D&cauthor=true&cauthor_uid=18436948), [Andrew D Oxman](https://www.ncbi.nlm.nih.gov/pubmed/?term=Oxman%20AD%5BAuthor%5D&cauthor=true&cauthor_uid=18436948),  [Gunn E Vist](https://www.ncbi.nlm.nih.gov/pubmed/?term=Vist%20GE%5BAuthor%5D&cauthor=true&cauthor_uid=18436948),  [Regina Kunz](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kunz%20R%5BAuthor%5D&cauthor=true&cauthor_uid=18436948),  [Yngve Falck-Ytter](https://www.ncbi.nlm.nih.gov/pubmed/?term=Falck-Ytter%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=18436948)
3. Preferred Reporting Items for Systematic Reviews and Meta-Analyses by David Moher , Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman
4. Effect of Online Reviews on Consumer Purchase Behavior by [Zan Mo](https://www.researchgate.net/scientific-contributions/Zan-Mo-2080192504), [Yan-Fei Li](https://www.researchgate.net/scientific-contributions/Yan-Fei-Li-2080196567), [Peng Fan](https://www.researchgate.net/scientific-contributions/Peng-Fan-2080187533)
5. Review rating prediction using combined latent topics and associated sentiments: an empirical review by [Anbazhagan Mahadevan](https://link.springer.com/article/10.1007/s11761-019-00278-6#auth-Anbazhagan-Mahadevan), [Michael Arock](https://link.springer.com/article/10.1007/s11761-019-00278-6#auth-Michael-Arock)

* Motivation for the Problem Undertaken

Understanding how people make online purchasing decisions is of growing importance. The goal of the present research is to contribute to understanding how comments are classified as ratings. A numeric rating and its accompanying text review is the most widely available form of user feedback. A measure which encapsulates the contents of such reviews is often necessary as they have been found to significantly influence the shopping behavior of users.

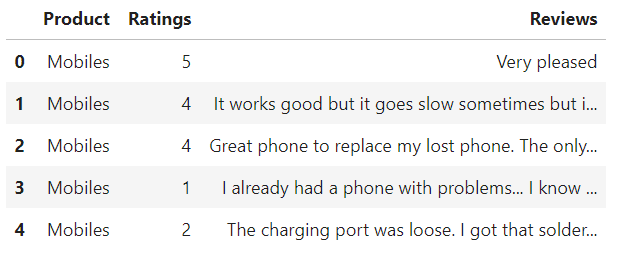
The review rating prediction tries to predict a rating corresponding to the given review. An approach that performs review rating prediction task by using the latent topics extracted from reviews and their associated sentiments. The main motivation of this project is rating based on the reviews.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

[Natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP) is a field that focuses on making natural human language usable by computer programs. **NLTK**, or [Natural Language Toolkit](https://www.nltk.org/), is a Python package that you can use for NLP. A lot of the data that could be analyzing is [unstructured data](https://en.wikipedia.org/wiki/Unstructured_data) and contains human-readable text. Before analyzing that data programmatically, first need to preprocess it. Various statistical, mathematical, analytical algorithms are used. Experimental design, etc are done for the problem.

* Data Sources and their formats



The data is received from the client. The data is in excel format which can be imported using pandas from local library easily.

* Data Preprocessing Done

Checked for null values where there is only 3 null values present. The nullvalues are removed since the dataset is too large which will not have an impact. No correlation found and skewness found. Checking the balance of the dataset and convert all type of the columns from object tyope to string type. Checked for duplicates and removed it. Checked the length of strings. Convert all messages to lower case, replace email addresses with 'email', replace URLs with 'webaddress', replace money symbols with 'moneysymb', replace 10 digit phone numbers (formats include paranthesis, spaces, no spaces, dashes) with 'phonenumber', replace numbers with 'numbr'. Removed all the punctuations. After cleaning the dataset, tokenize the reviews column, lemmatize it and checked the length of string in each row. Transformed the data into vectors only which will be considered by the model and the target column and labels are separated.

* Data Inputs- Logic- Output Relationships

There is only 2 columns used to detect the ratings using the reviews. The features which are helping to detect the ratings is visualised using matplotlib and seaborn. The relationship between the features are determined. After data cleaning, pre-processing, model is built.

* State the set of assumptions (if any) related to the problem under consideration

In this age of the Internet, there has been a remarkable increase in reviewing the products and rating the products. The purchasing/buying behaviour of the people are changed based on the reviews and ratings. Only after checking out the model after building, we will know how reviews can be checked to tell the ratings based on the reviews text.

* Hardware and Software Requirements and Tools Used

Any laptop and computer can be used as hardware. Processor used is Intel(R) Core(TM) i7-4510U CPU. System type is 64-bit operating system, x-64 based processor. RAM of the systerm is 8.00 GB. Microsoft Windows 8.1 version 6.3 is the OS used. Python 2.7.10 is used with the interface Jupyter notebook with many installed libraries.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

There are statistical and analytical approaches in problem-solving. Data cleaning, data pre-processing, comes under statistical approach whereas data manipulation, vectorizing, creating ML models comes under analytical approach.

**Statistical:**

* + - ***Experimental Design*** - Methods to design systematic experiments to compare the effect of independent variables on an outcome
  + Univariate - measure relies on only one variable - Statistical summary, information on variables, etc
  + Bivariate - measure relies on two variables - Correlation, multicollinearity
    - ***word\_tokenize -*** used to tokenize the words. i,e., separate the words using comma
    - ***WordNetLemmatizer*** - used to convert the similar words to a meaning word which is equally and similar to it
    - ***stopwords*** - Used to detect the words which does not help in determining the language or determining the result of the model
    - ***Re-sampling Methods*** - Train Test Split is used to systematically split a dataset into subsets for the purposes of training and evaluating this predictive model
    - ***Statistical Hypothesis Tests*** - Cross validation quantifies the likelihood of observing the result given an assumption or expectation about the result whether the model is overfitting/underfitting or fitting good
    - ***Estimation Statistics*** - GridSearchCV is used to quantify the best parameter from the listed to fit in the model and give better result. It uses data analysis framework which has a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results

**Analytical:**

It concerns the design and development of algorithms.

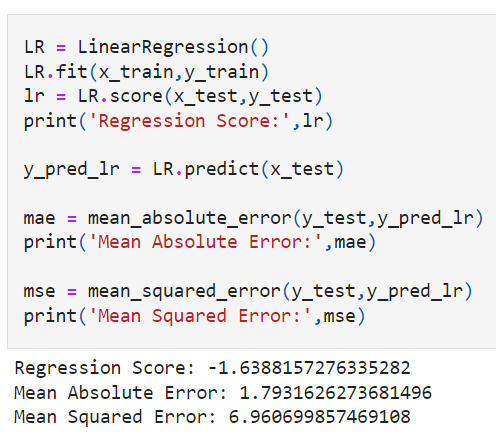
* ***TfidfVectorizer*** - Convert all the alphabets to vectors to make sure the model can understand the language
* Testing of Identified Approaches (Algorithms)

Various evaluation metrics can be used for this classification type of model. Some of the popular algorithms are the following:

* Linear Regression
* k-Nearest Neighbors
* Decision Trees
* Support Vector Regressor
* Random Forest Regression
* Run and Evaluate selected models

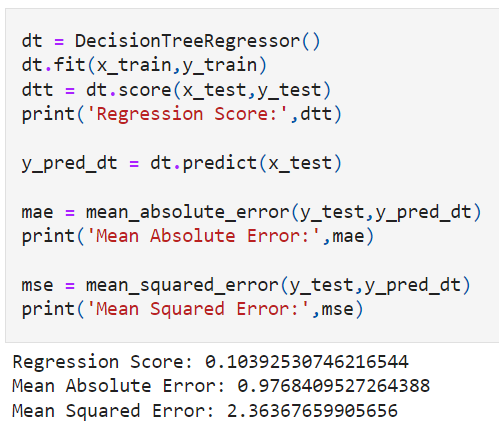
Regression Score - Percentage of correct predictions for dataset

***Linear Regression:***



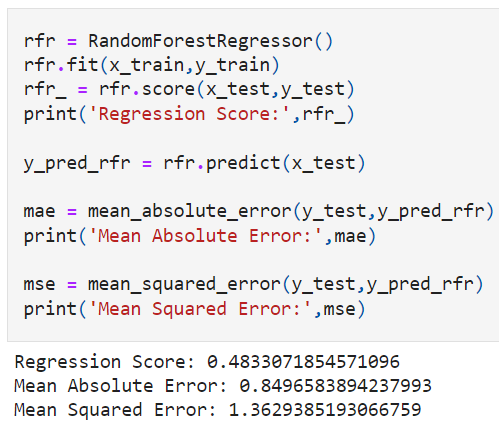
Regression Score - -1.6

***Decision Tree:***



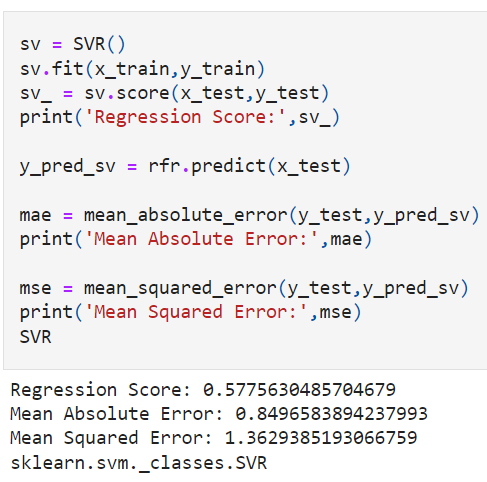
Regression Score - 0.10

***Random Forest Regressor:***



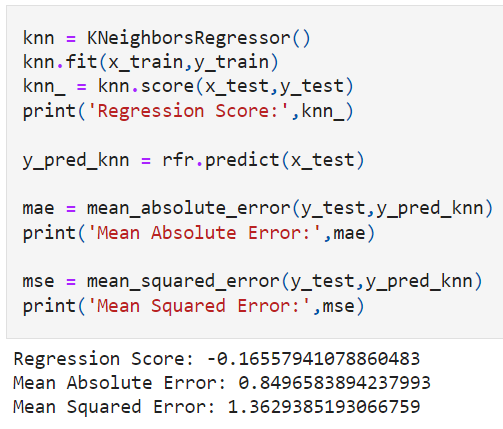
Regression Score - 0.48

***Support Vector Regression:***



Regression Score - 0.57

***KNeighbors Regressor:***



Regression Score - -0.16

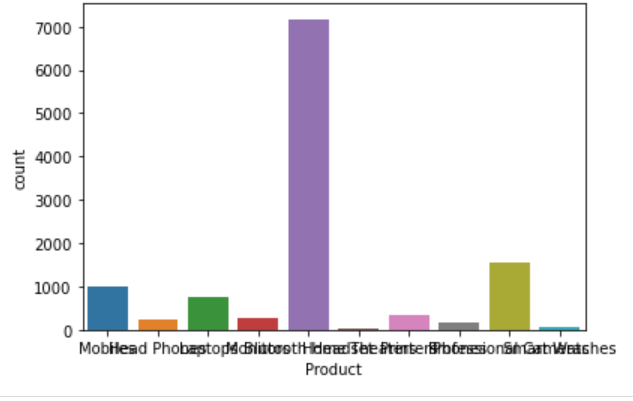
* Key Metrics for success in solving problem under consideration



|  |  |
| --- | --- |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| WordNetLemmatizer | To convert the similar to a unique word having the same meaning in natural language |
| word\_tokenize | It is used to separate the sentences into words |
| stopwords | Used to check the words which does not help in determining the solution for the model |
| string | This package is used to remove the punctuations, numbers, etc |
| R2 score, mean\_squared\_error, mean\_absolute\_error report | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Linear Regression, DecisionTreeRegressor, RandomForestRegressor, SVR, KNeighborsRegressor | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| GridSearchCV | For hyper parameter tuning |
| TfidfVectorizer | To convert the natural language to vectors only which can be read and analysed by the models |
| Plot\_roc\_curve | To check whether the model is good by checking the area under the curve |

* Visualizations

Matplotlib and seaborn is used for visualisations



Visualised the count of variables used in the model.

* Interpretation of the Results
* The count of variables is visualized to understand how the data is balanced.
* In data preprocessing, duplicates are removed which helps to get unbiased result
* The null values are removed since it was in small amount
* The sentences are tokenized to words
* The words are lemmatised instead of stemming which will help us to understand the meaning of the word
* Stopwords are removed, so that the model can get a good unbiased result

**CONCLUSION**

* Key Findings and Conclusions of the Study

The relation between the input and output variables cannot be found her because we are doing a natural language processing model using natural language toolkit in python. Correlation is checked and nothing is found. The features are consolidated to find whether the comments are malignant or not to get better results. From the models, support vector regressor gives better result and also on comparing the cross validation the model fitted well.

Check the hyper parameter tuning and use the better attributes. Save the model and that can be used for predictions later since the model is trained well.

* Learning Outcomes of the Study in respect of Data Science

Visualising the features tells the balancing of the dataset. After checking the correlation, nothing is found. The data has only 3 null values which is removed and the data is pre-processed before building the model since we are building a natural language processing model.

Support Vector Regressor is good to go with the model. Random forest regression and SVR takes time for training. SVR takes very long time to train the model. Cross-validation took time to complete and checked whether the model is underfitting and overfitting the model.

Support Vector Regressor gave 57% accuracy approximately. Even after hyper parameter tuning, the model score did not increase. So, we can use Support Vector Regressor for further predictions.

* Limitations of this work and Scope for Future Work

Data is clearly explained. Training data and testing data is separated using train test split. The model can be tested by using stemming during pre-processing.