

Rating Prediction Project

Submitted by:

Ruchita Parmar

**ACKNOWLEDGMENT**

Abstract E-commerce industry is not only the industry or a centre of online shopping, rather it is now the centre of global business. All over the world is excited about product reviews, rating and popularity etc. A huge data is available online about these online E-commerce websites. We have used the Flipkart and their product rating from the available website to create our dataset. The machine learning classification algorithms are applied to the dataset. Lastly an efficient model is developed to predict the available product rating based on available reviews. The model gives a good classification measures with the data set. Users post online reviews for various business product and services. Review Rating Prediction is a task of predicting users' star rating based on the text review of the user. In this project, we treat Review Rating Prediction as a multi-class classification problem, and build several different prediction models by combining feature extraction methods based on different machine learning methods 1) Logistic Regression, (2) Naive Bayes, (3) SVM etc.

**INTRODUCTION**

The rise in online shopping has brought a significant rise in the importance of textual customer reviews. There are thousands of review sites online and massive amounts of reviews for each and every product. Nowadays customers have changed their way of shopping. 60-70 present of customers say that they use rating filters to filter out low rated items in their searches.

The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is important to companies that support these reviews, companies like Flipkart and Amazon.

The customer will make a decision to buy a product if he or she sees valuable reviews posted by others, especially the user’s trusted friend. We believe reviews and reviewers will do help to the rating prediction based on the idea that 5-star ratings may greatly be attached with extremely good reviews. It’s also agreed that different people may have different sentimental expression preferences. For example, some users prefer to use “good” to describe an “excellent” product, while others may prefer to use “good” to describe a “just so so” product. User’s rating information is not always available on many review websites.

* Business Problem Framing

This project focuses on two approaches to predict review ratings. Sentiment analysis studies the textual contents of existing reviews and learn positive or negative opinions. The recommender predicts ratings based on user-business pairs and recommends to a user the unreviewed businesses with highest predicted rating.

* Conceptual Background of the Domain Problem

Crowd-sourced online review sites like Flipkart collect opinions of businesses and provide recommendations for users. Every minute 26380 reviews are posted by Yelp users. And 80% of users visit Yelp with the intention to make a purchase based on reviews.

* Review of Literature

The dataset used for this task is scrape from Flipkart (http://www.Flipkart.com). Most of the predictive tasks previously performed on this dataset have rating predictions primarily based on user and business attributes. However, research has been carried out, not just in the general area of text mining and sentiment analysis, but in text mining for predictive tasks in review and rating systems. The impact of text derived information has been previously studied at the sentence level, with the help of the topic information on various datasets. Various methods have been adopted in the past, including regression, bag of opinions method and classification .In movie reviews, it has been observed that Naive Bayes’ had a slightly better accuracy than the SVM method. However, this was in combination with other features of the dataset. Hence, the results differ from the ones chosen here.

Motivation for the Problem Undertaken

Data analytics helps e-commerce players like Flipkart to identify both loyal and new customers by using data extraction and segmentation for tracking browsing habits and spending patterns. Key Metrics for success in solving problem under consideration. What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

Metrics are used to monitor and measure the performance of a model (during training and testing), and don’t need to be differentiable.

I noticed that online advertisements of a product spring up after you have searched for it on an e-commerce website? Such targeted advertising, meant to remind you of unfinished business, is a result of high-end data analytics, machine learning and complex algorithms to bolster sales. Flipkart and other e-commerce players keep track of every click you make on their portals and then predict what you are most likely to buy from them with the help of review and ratings.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Every classification algorithm is built up with strong mathematical models and logic. Though the decision tree is one of the classification algorithms every algorithm is bound to make one decision before starting of making the trained model. The logic of the SVM is for finding out the optimal separating hyper-plane that maximizes the margin of training data. There is a key challenge to find out the appropriate algorithm for each application in machine learning. The application depends on training time, prediction speed and accuracy, number of parameters, number of features, validation strategies, estimation of uncertainty condition, etc. Hence, in machine learning, the amount of math like probability, calculus, linear algebra, complex algorithm, etc. is used to justify the algorithm. The usage of algebra or linear algebra for Eigen value calculation, principal component analysis and probability is used for prediction and choice of the classes as this is a critical part in machine learning algorithm. Markov chain model is a statistical and mathematical setup which has some hidden layers and can be represented as the simple Basyian network which is directly visible to the observer. This model has a remarkable contribution in the field of supervision, reinforcement of learning and for pattern recognition.

* Data Sources and their formats

Data source were used from Flipkart an online e-commerce website and I scraped Different-Different type product like mobile, smart TV , headphone, home appliances, Books, health and Wellness, etc.

* Data Preprocessing Done

We have performed the following preprocessing on the data:

* **Removed punctuations**:- In this step, all the unwanted character which are not required by the classifier or which do not contribute in making a review positive or negative will be removed and the only alphabet will be left over, which will be in both upper case as well as of lower case.
* **Removed the stop words**:- Stop words are those words which do not contribute to making a particular sentence positive or negative. Example words like ‘is’, ‘a’, ‘the’, etc. are considered as stop words. Stop words can be in positive as well as in negative reviews and it does not affect the polarity of the review.
* **Tokenization: -** Tokenization is taking a sentence into consideration and breaking it up into its individual words or tokens and tokens are mostly a single word. The use of tokenization is in Bag of Words (BoW) model in which every column will represent different words.
* **Stemming and lemmatization**:- Stemming is the process of reducing a word to its word stem that affixes to the roots of words known as a lemma. Stemming is important in natural language processing (NLP) and text categorization.

After all these data pre-processing steps the pre-processed data is stored in a list which is called a corpus. Corpus means a collection of related written text

**TF-IDF**:- In TF-IDF some semantic information is preserved as uncommon words are given more importance than common words Here the word ‘awesome’ in the sentence ‘You are an awesome guy’ will get more important. TF-IDF model give more importance to specific, uncommon and important words.

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

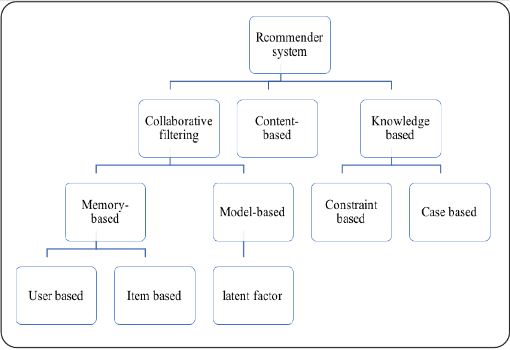
We followed 2 main problem-solving approaches, which is listed as below.

**Sentimental Approach** – In Sentiment analysis or opinion mining, we analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes, where the term opinion relies on the attitude and associated information by the person who wrote this opinion whereas Sentiment is the underlying positive or negative feeling implied by opinion. This approach further gets divided in 3 different approaches.

1. Lexicon based – In this lexicon, each word has a polarity value such as 1 for positive and 0 or -1 for the negative. And from that, the polarity of a sentence can be calculated.
2. Corpus based – One of the machine learning algorithms is used such as Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN) where a manually an-notated corpus is used in the training process of the classifier. After building the model, this model will be used in the classification process of the testing dataset.
3. Hybrid based – This approach is the combination of both the above approaches i.e Lexicon + Corpus.

**Recommender Approach** - Recommender systems are filtering systems that show specific information to the user or customer and attempt to predict a rating or products. Most of the Companies or websites with different areas allow the user to provide a feedback about what did they like and dislike. The most famous and oldest way of feedback is ratings another kind is the textual notes or comments. The main idea of recommender systems is recommending an entity that belongs to the user interests. Usually, the recommendation process depends on the user and items interactions, past behaviour, relation to other users, item similarity or it can be a knowledge-based recommender system that relies on the user requirement rather than depending on the user past history. There are three methods for the Recommender as listed below and shown in figure:

1. Collaborative filtering
2. Content-based method
3. Knowledge-based method



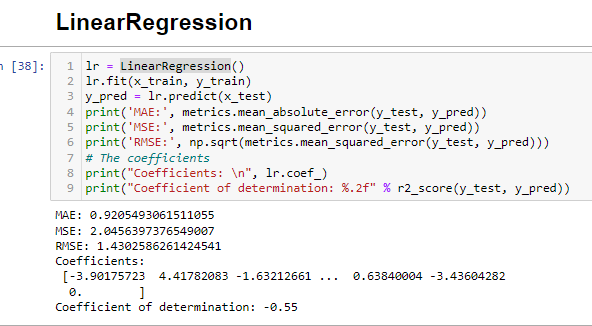
Recommender system methods

* Testing of Identified Approaches (Algorithms)

The possible problem for a machine learning algorithm to be able to solve are vast in terms of area wise applications and scope for future research. There are two ways of handling such problems. First is test away i.e Test all possible algorithms on your data to see which works best for you. There are both pros and cons to this approach. The pros would be that you would definitely know that one algorithm or a set of algorithms are better choices for your problem statement. However, the second way which is to try to understand what the algorithm does before deciding if it is a good fit for your problem or not. So, here we were going with only 5 listed below Algorithms, used for the training and testing.

1. **Linear Regression**
2. **Logistic Regression**
3. **Decision Tree**
4. **KNN**
5. **MultinomialNB**

* Run and Evaluate selected models
* **Linear Regression** – It’s a linear approach to modelling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). It is about learning the relationship between the dependent and independent variables from a pile of historical data. From a mathematical point of view, linear regression is about fitting data to minimize the sum of residual between each point and the predicted value.



* **Logistic Regression** – The logistic regression statistic modelling technique is used when we have a binary outcome variable. Below is the function for logistic regression:



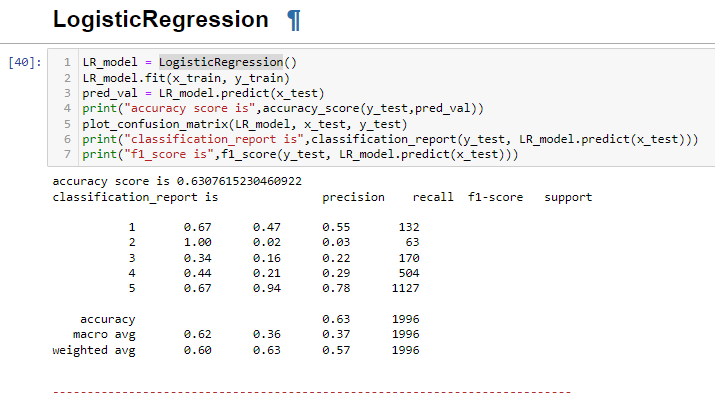
Where E is log base

X is the numerical value that needs to be transformed.

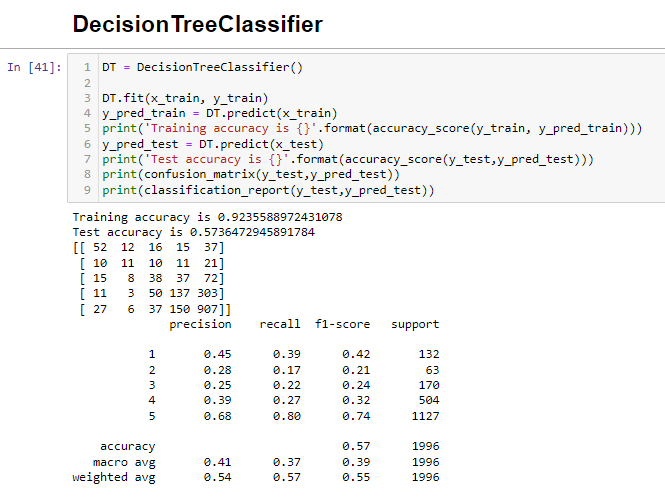
If we feed an output value to the sigmoid function, it will return the probability of the outcome between 0 and 1. If the value is below 0.5, then the output is return as No/Fail/Deceased. If the value is above 0.5, then the output is returned as Yes or Pass or Deceased.

Assumptions of Logistic regression:

* Independent variables show a linear relationship with the log of output variables.
* Non-Collinearity between independent variables. That is, independent variables are independent of each other.
* Output variable is binary.



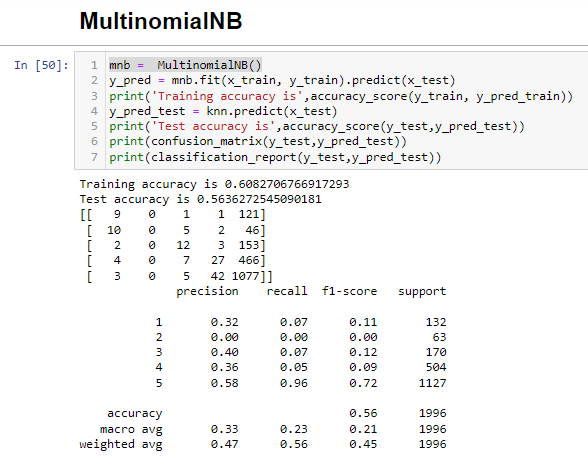
* **Decision Tree** – Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. There are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.



* **KNN** – The k-nearest neighbour’s algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While this is technically considered “plurality voting”, the term, “majority vote” is more commonly used in literature. The distinction between these terminologies is that “majority voting” technically requires a majority of greater than 50%, which primarily works when there are only two categories. When we have multiple classes—e.g. four categories, we don’t necessarily need 50% of the vote to make a conclusion about a class; we could assign a class label with a vote of greater than 25%.



* **Multinomial NB** – Multinomil Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.



* Visualizations

For visualizing our model we only go through historical 5 W’s.

Why, Who, What, When and Where.

**Why** we need to understand the model’s decision-making process. The extent of this problem becomes especially clear in the case of rating prediction. Building machine learning models is an iterative process full of experimentation. Finding the optimal combination of hyper-parameters can be quite challenging. Visualization can accelerate this process. Choosing the best model out of an ensemble of well-performing models can be simply reduced to visualizing parts of the model which offer the highest accuracy or lowest loss while ensuring the model doesn’t over-fit.

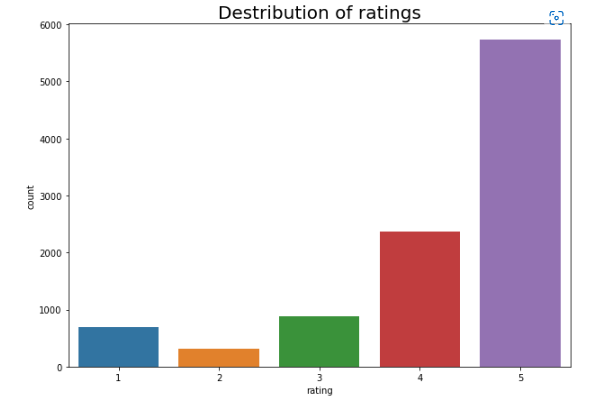
**Who** mainly focus on developing, experimenting with, and deploying models are the ones who will benefit the most from visualization.

**What** - The first and main thing we can visualize is the model architecture. This tells us how many layers there are, in what order they’re positioned, and more. This also includes the computational graph that defines how a model would train, test, save to the disk and checkpoint after epoch iterations.

**When** - Using visualization while training is a good way to monitor and track model performance. There are techniques like attribution visualization, to regenerate an image with important regions highlighted, and feature visualization, to generate an entirely new image that supposedly is representative of the same class.

**Where** - Visualization has been used a lot in domains like autonomous driving, urban planning, medical imaging to increase user trust in the models. Combining visualization with research has led to the creation of tools and frameworks for model interpretability and democratization.

Rating distribution.



**CONCLUSION AND FURTHER INTERESTS**

This paper explores two ways in which online reviews influence the consumers’ purchase behaviours. Consumers search the business of interest, read posted reviews and perceive sentiments to form their judgements on business performance. The review websites predict how users would rate different businesses and recommend to consumers the businesses they might rate high. Thus, sentiment analysis on a per review basis and recommender building based on user-business pairs are main approaches to rating predictions in this paper. Natural Language Processing stimulates the human-level understandings of the text reviews. In sentiment analysis, important text features are extracted with vectorizers and the encoded documents are used to train binary classifiers to get positive (rating star of 4 or 5) or negative (rating star below 4) rating predictions. TfidfVectorizer is a better vectorizer in most cases. As for classification algorithms, logistic regression classifier (Acc =0.63, 0.918). The encoded documents can also be used to find top words for positive or negative predictions and search for similar reviews across documents.

The neighbourhood collaborative filtering recommender built in this paper gives user-specific rating prediction. It predicts five-star rating by nearest item neighbors and common reviewer’s ratings. Ranking predictions can give top recommendations for a user. Testing on the whole dataset, the built recommender gives decent accuracy measures (Accuracy = 0.63627, RMSE = 0.71174). Although sentiment analysis and the recommender system explores reviews from different perspectives, the comparison scan still be made to provide some hints. Binary sentiment analysis is on a per review basis and gives seemly higher accuracy measures .And it also adds on real-world interpretability of review ratings by learning from word features. However, the models can only tell positive or negative sentiments in reviews instead of five-star rating estimates. The neighborhood-based recommender is modeled to get the five-star predictions specific to a user and a business. The predictions can be directly compared to actual ratings with decent accuracy measures. And the user-specific recommendations can be made.

Thank you