PSTAT 235 Final Project Report

Amazon Reviews:
Unlocked Mobile Phones

By: Hongshan Lin, Zijie Fang, Kexin Zhou, Eileen Zhu

Table of Contents

1. Abstract

In this project, we mine patterns in around 400k Amazon phone reviews to predict ratings of items. In order to address the problem effectively, we divided the problem into the following modules: missing values and categorical data handling, data preprocessing and feature extraction, and model training and evaluation. We developed algorithms to take care of data preprocessing and trained multiple different models. To make it more solvable,we used many graph in the Exploratory Analysis: we drew the world cloud to show the top frequent words in reviews, and plotted the distribution of rating and brand. After feature extraction, we explored a few different training models, including Decision Tree, Naives Bayes, Random Forest, Logistic Regression and Lasso Regression. We find that Decision Tree yields the best performance with the highest accuracy of over 54% ,highest F1 score and highest weighted recall.

2. Data and Methods

2.1 Data Source

Amazon Reviews: Unlocked Mobile Phones
https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones

2.2 Background

Mobile phones have revolutionized the way we purchase products online, making all the information available at our fingertips. As the access to information becomes easier, more and more consumers will seek product information from other consumers apart from the information provided by the seller. Reviews and ratings submitted by consumers are examples of such type of information and they have already become an integral part of customer's buying-decision process. The review and ratings platform provided by eCommerce players creates transparent system for consumers to take informed decision and feel confident about it.

Therefore, we decide to choose a text-mining analysis Amazon.com is a treasure trove of product reviews and their review system is accessible across all channels presenting reviews in an easy-to-use format. The product reviewer submits a rating on a scale of 1 to 5 and provides own viewpoint according to the whole experience. The mean value is calculated from all the ratings to arrive at the final product rating. Others can also mark yes or no to a review depending on its helpfulness – adding credibility to the review and reviewer. In this study, we analysed more than 400 thousand reviews of unlocked mobile phones sold on Amazon.com to find insights with respect to reviews, ratings, price and their relationships.

2.3 Data Description

The original dataset consists six columns: Product Title, Brand Name, Price, Rating, Reviews, Review Votes. There are more than 400,000 reviews covering close to 4,400 unlocked mobile phones. Nevertheless, the dataset we used in this project is a subset of the original dataset. The subset is randomly selected and consists of 100,000 reviews and includes three columns: Brand Name, Rating, and Reviews.

2.4 Research Question

This statistical analysis had the following goals: Perform exploratory analysis of mobile phones ratings and reviews from Amazon website; Predict the expected rating based on the reviews; Find the best model representing the most accurate prediction.

2.5 Data Preprocessing

After getting the random sampled dataset, first of all, we removed missing values from all three columns. It is important to handle missing values properly in order to successfully manage data. Due to the large dataset, we chose to remove missing values. By cleaning the missing values, we are able to perform analysis correctly and draw accurate inference about the data.

summary	BrandName	Rating	Reviews
count	84391	84391	84391
mean	null	3.8177649275396663	null
stddev	null	1.5459023383844865	null
min	"BlackBerry Storm...	1	!!!!!Update!!!!! ...
max	worryfree	5	🤗

Figure 1: Data Frame after Removing Missing Values

Next, in order to prepare for text mining, we translated reviews into individuals words using RegexTokenizer, including remove any punctuation. Then we filtered the words by removing stop words, such as the, a, I, etc.using StopWordsRemover. We translated the filtered words into features using TF-IDF and Ratings into Labels using StringIndexer to prepare for model building.

2.6 Exploratory Data Analysis

2.6.1 General Rating Analysis

Since we have already transformed Ratings into Labels in the data preprocessing, we just simply counted the total number of each Label and drew a Label Distribution (Figure 2) to represent the mobile phones rating distribution. Label 0.0 is actually the rating 5.0 and it is obvious that most of Amazon mobile phones reviews got 5-star ratings.

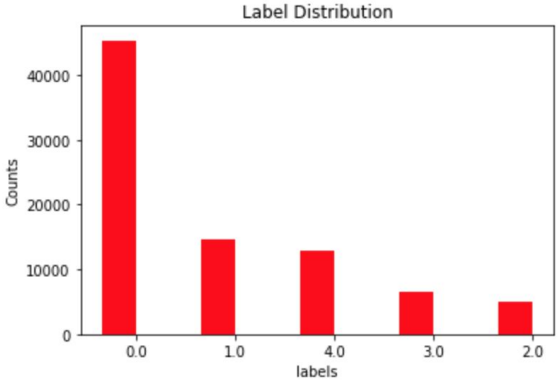

Figure 2: Label Distribution

Also, we are curious about the relationship between the lengths of reviews and ratings. On the Figure 3 below, you can see the higher rating tends to have longer reviews.

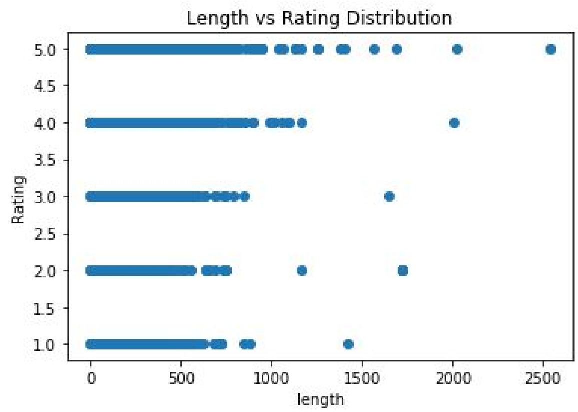

Figure 3: Length vs. Rating Distribution

2.6.2 General Reviews Analysis

In order to analyze Reviews, we selected only tokenized and stop words removed reviews from the dataframe. Then, we transformed that selected dataframe to pandas and saved it as a TXT file. We used two ways to show word frequencies. First, we imported the TXT file as a RDD and created the Frequency Distribution for Reviews which exactly shows how many times a certain word showed up on all the reviews in the descending order. Second, we imported the TXT file as String and created the Word Cloud for Reviews. The Word Cloud shows the most frequently used words in reviews by the size of words. When a word shows up more, it would have a bigger size. In Figure 4, words like "great", "phone", and "problem" have the biggest sizes, which means they show up the most times in reviews.

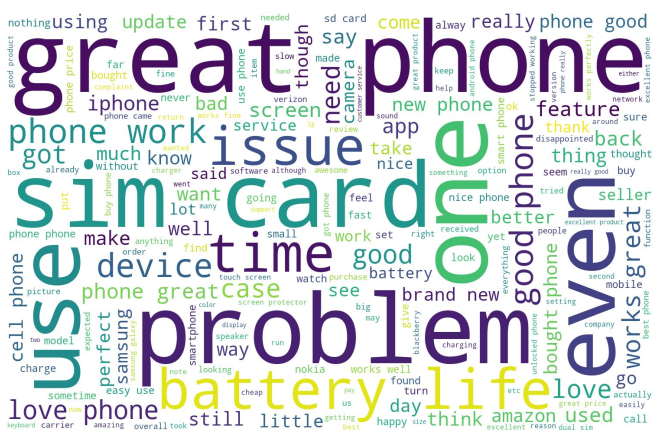

Figure 4: Word Cloud for the Most Frequently Used Words in Reviews

2.6.3 Sentiment Words Analysis

From the Word Cloud for the Most Frequently Used Words in Reviews, we found words like "great", "good", and "love" could be considered as positive words, while words like "problem", "bad", and "disappointed" could be considered as negative words.

In addition, in order to find more sentiment words in reviews, we applied Word2Vec to produce word embeddings to find synonyms of words "good" and "bad" by computing cosine distances. As a result, the top 5 synonyms of "good" are "nice", "great", "decent", "fantastic", and "solid"; the top 5 synonyms of "bad" are "good", "disappointing", "bad", "terrible", and "refurbished". Even though we found the positive word "good" always shows up with the negative word "bad", we still regard "good" as a positive word. Then, we created two sentiment words dictionaries for positive words and negative words from the general reviews analysis and Word2Vec application. Based on the word counts in Figure 5, the frequencies of positive words are much higher than negative words, which means more reviews are positive. In future studies, we will talk about more potential sentiment words analysis.

	wordCount		wordCount
great	21310	not	7712
good	20170	problem	7228
like	13156	issue	4387
love	11966	never	3089
perfect	6033	bad	2952
nice	5947	slow	2063
recommend	4332	disappointed	1601
best	3831	complaint	1053
thank	3538	returned	1032
amazing	2437	terrible	681

Figure 5: Frequency Distribution for the Most Frequently Used Positive/Negative Words in Reviews

2.6.4 Brand Rating Analysis

In the process of analyzing the brand reputation, we counted the number of 5-star ratings for each brand and computed the average rating for each brand. Figure 6 shows that Samsung has the greatest number of 5-star ratings, so we decided to do a samsung review analysis in next part.

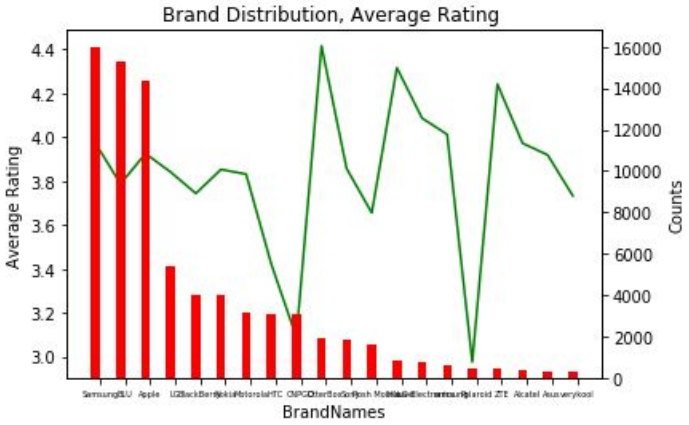

Figure 6: Brand Distribution & Average Rating
(Redbars = 5-star counts, Green line = average rating)

However, the 5-star counts for each brand cannot accurately represent the brand reputation. It is because the more mobile phones sold, the more number of 5-star ratings would be found. To be more representative, we calculated the 5-Star Rate by dividing the number of 5-star ratings by the total number of ratings for each brand. As the word cloud shows, the brand Getnord has the highest 5-star rate and the brand Seawolf Technologies has the second highest 5-star rate.

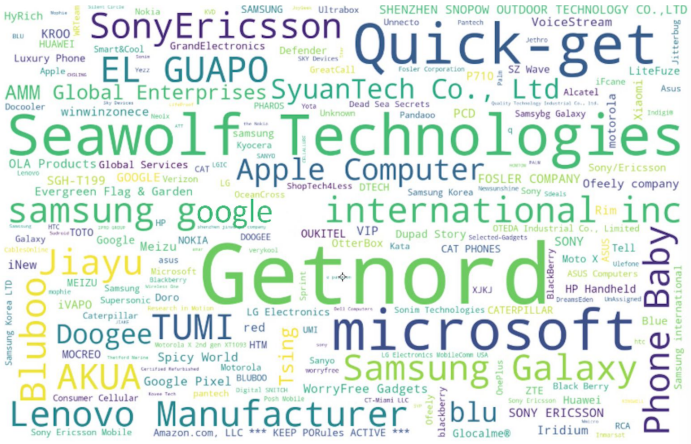

Figure 7: Word Cloud for Brands Based on 5 Star Rating
(5-Star Rate = number of 5-Star Rating / Total number of Rating)

2.6.5 Samsung Reviews Analysis

Since Samsung is the best seller of mobile phones at Amazon website, we did reviews analysis only for the brand Samsung. First, we selected only samsung ratings and reviews, then we tokenized words by pipeline, transformed the selected dataframe to pandas, and saved it as a TXT file to import as a string. Similarly than previous analysis, we created the word cloud for the most frequently used words in Samsung reviews (Figure 7). Obviously, the 3 largest words from the word cloud are “great”, “good”, and “love” which is telling us that the reviews of Samsung are really good. Thus, if someone is asking us for a mobile phone recommendation, just based on this word cloud, we could recommend Samsung.

Figure 7:Word Cloud for the Most Frequently Used Words in Samsung Reviews

2.7 Models

2.7.1 Split Data

After preprocessing the data, we get a dataset with label being the rate and features being the vector that is transformed from reviews. Our goal is to find models that properly predicts ratings from reviews, so the next step is to build models and select the best model for prediction. The data we use for building models is a subset of the preprocessed data,training data, and the rest-part will be used to test accuracy, test data. The training data will be 70 percent of the whole dataset, while the test data be 30 percent of the total data.

As our predicted value, rating, can either be considered as label or quantitative integer, we use both classification and regression models to predict. The classification models we use are decision tree model, random forest model and naive bayes classification model, while we also tried logistic regression and lasso regression to see their performance.

2.7.2 Decision Tree（Landmark Model）

Decision tree is one of the first models we think of to fit our data. Decision tree is good at transparency. The result is derived from making decision of true or false on each node. It can handle multiclass classification and is suitable to be our benchmark model. To evaluate the model, we calculated test error, f1 score, weighted precision and weighted recall for each model. The test error for decision tree model is 0.4517, f1 score be 0.4371,weighted precision be 0.3738, and weighted recall is 0.5483.

2.7.3 Random Forest

Random forest is the combination of several decision trees. As each decision tree give out different predicted result, random forest outputs the most frequent result. Random forest is also good at transparency and usually performs well. The only downside for fitting our data into random forest is that random forest requires each tree to be independent to each other, but our predictor which is extracted from reviews has some correlation between each other. We fit a random forest with 10 number of trees and tunes the parameter numTree and maxDepth using grid search cross validation. Tuning parameters will help us get the best random forest model,and to evaluate the model, we calculate the same statistics as decision tree. The test error for random forest model is 0.4595, f1 score be 0.3841,weighted precision be 0.3873, and weighted recall is 0.5405.

2.7.4 Naive Bayes

Naive Bayes model is based on Bayes Theorem, which allow us to convert P(A|B) to P(B|A). On our dataset, the naive bayes model will return the predicted rating that has the highest probability based on it's review. Naive bayes models is popular in text classification because it is easy to understand and usually have higher accuracy than other models. We first fit a naive bayes model with smooth equal to 1, and then we tune parameter smooth using k-fold. The result shows that

we get the best model when smooth is 1. The evaluation statistics for naive bayes are test error be 0.4585, f1 be 0.4038, precision be 0.4096 and recall is 0.5416.

2.7.5 Logistic Regression

Logistic regression is based on bernoulli distribution and returns the max likelihood model possible. The logistic regression model is more abstract to understand when comparing with other models, but we are interested to build a more statistical based model to see how it performs. We get the evaluation statistics for logistic regression model to be test error 0.4650, f1 0.3805, precision be 0.3731 and recall is 0.5350.

2.7.6 Lasso Regression

Lasso regression is helpful as we can extract coefficients from lasso model to see which predictor is important and which is not. However, our lasso regression model returns all the coefficients to be 0, thus lasso regression may not be a good model for our data.

3. Results

3.1 Model Evaluation

	Test Error	F1 Score	Weighted Precision	Weighted Recall
Random Forest	0.459491	0.384123	0.387321	0.540509
Decision Tree	0.451742	0.437111	0.373766	0.548258
Naive Bayes	0.458458	0.403813	0.409568	0.541542

Figure 8:Model Evaluation Results

This chart shows all the evaluation statistics of three classification models and logistic regression model after cross-validation. Decision Tree have the highest accuracy of over 54% ,highest F1 score of 0.4371 and highest weighted recall of 0.5483. It shows that Decision Tree which the landmark model is also the champion model.

4. Conclusion

In this Project, our team mine patterns in more than 400,000 reviews of unlocked mobile phones sold on Amazon.com to predict ratings of items.We have worked out a plan to solve this problem through the following phases: data preprocessing, feature extraction and variable construction, feature scaling, missing values and categorical data handling, model training and model evaluation.We build both Regression and Classification models which include Random Forest, Decision Tree, and Naive Bayes to predict the expected rating based on the reviews. Base on the ROC score and cross validation we concluded our champion model is Decision Tree.

4.1 Future Studies

Although we conclude our best model is Decision Tree, in fact all the 3 models do not perform very well to predict the rating, and the highest precision among our models is only around 0.5.

Reaching a high precision can be a challenge in the cases. Therefore on the future study we aim to improve the model performance. First, we want to enlarge our sample size from 100k to 500k to get a more comprehensive analysis. Second, we want to extract a better feautre by normalization and discrimination. Third, we want to try some more advanced models such as logistic regression, Support Vector Machines, and recurrent neural network to fit our data. Finally we will tune parameters in the decision tree, random forest model to prevent overfitting.

In addition ,through this project we find the natural language process, an attempt to make a computer understand human language. is interesting and valuable in the business field. The big picture of our project is to continue study of the amazon reviews in NLP. For instance, we want to use sentiment analysis to prevent the abuse and fake review in both 5-star and 1-star review. Sometimes, sellers buy blocks of fake Amazon customer accounts by the thousand. Then they use people or software bots to write fake five-star reviews for their own products to attract customers. They're careful to make subtle changes to each review for example — so that Amazon's algorithms won't spot the duplicates and vice versa for the negative reviews. If we are able to solve this problem, the amazon review will be more reliable.

Figure 9:An Example of Fake 5-Star Reviews