Data Analytics 350 Amazon Lab

Sally Hyde and Matthew Rinker

***Abstract*—In this report, we explore a dataset from Amazon and attempt to predict review price from a variety of predictor variables. Our analysis is based upon feature engineering (particularly sentiment analyses) and linear regression.**

I. INTRODUCTION

In this lab, we explored a dataset containing Amazon's fine food reviews, which included 586,454 reviews from a period of 10 years. The data also included the review Score, which is the ranking from 1-5 stars that reviewers ranked the food item. We were tasked with building a model that could predict review score based on a limited number of fields we had. Using feature engineering techniques, namely sentiment analysis of the review text data, we were able to adequately predict review scores with a high R^2 value of 0.396, meaning that nearly 40% of the variation in the review scores data could be explained by our model. The sentiment analysis of the Summary and Text columns had significant effects on the score as we expected, and interestingly, the length of the Summary and Text columns also has significant effects.

II. METHODS

The original Amazon reviews data included 12 variables:

- **Id** : a unique identifier for each observation in reviews dataset
- **ProductId** : a unique identifier for each reviewed product
- **UserId** : a unique identifier for each user who left a review
- **ProfileName** : a the profile name of the Amazon customer leaving a review
- **HelpfulnessNumerator** : Total number of people who found review helpful
- **HelpfulnessDenomerator** : Total number of people who found review helpful OR not helpful
- **Score** : Main predictor variable; review score (1-5)
- **Time** : Data of review in terms of seconds
- **Summary** : Summary of review left by user
- **Text** : Main body of the review

Our primary strategy for this lab was to do a sentiment analysis of the text data (**Text** and **Summary**) for the Amazon reviews. Our rationale for this strategy is that reviews with positive sentiments would have higher scores, while reviews with negative sentiments would have negative scores. Inspired by the research of Gezici et al.[1], we included text and summary length in our model, because the literature shows that other features of text data such as text length are important in classifying text as positive or negative. This may have something to do with the typical length of positive and negative words.

After executing a sentiment analysis on the **Summary** and **Text**, we used the compound positivity and negativity results for each field in our model. Additionally, we computed the length of each observation in the **Summary** and **Text** columns and included these new columns in our model. Finally, we used the **HelpfulnessNumerator** and **HelpfulnessDenominator** fields in our model. We used each of the above six fields to predict review **Score** in an OLS linear model. To go about model validation, we used the Train/Test split cross validation method.

To submit our results to the class Kaggle competition, We split our data by specific **Ids** that had to be submitted to Kaggle. Thus, with roughly 100,000 rows in the submission, and 500,000 rows in all the data, our train/test ratio was roughly 80/20. Our RMSE value (from Kaggle) was calculated using the test data, and came out to be 1.058.

III. RESULTS

Variables	Coefficients
Intercept	3.426
HelpfulnessNumerator	.1365
HelpfulnessDenominator	-.1302
compound_sum	1.214
compound_text	1.002
SummaryLength	-0.005
TextLength	-.0002

Fig. 1. Coefficients for the Linear Regression Model (R^2 = 0.396)

The results of our model and cross validation can be seen in the below. From Fig. 1, it's clear that our model is an okay fit; with an R^2 value of 0.39, nearly 40% percent of the variance in the dependent variable (**Score**) can collectively be explained by the independent variables. Unsurprisingly, the two predictor variables with the largest

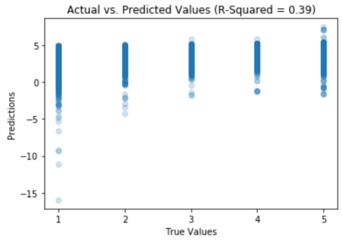

Fig. 2. Scatterplot of Actual vs. Predicted Values

effects are **compound_sum** and **compound_text**, meaning that our sentiment analysis produced valuable results.

References

[1] Gezici, Gizem & Yanikoglu, Berrin & Tapucu, Dilek & Saygin, Yucel. (2012). New Features for Sentiment Analysis: Do Sentences Matter?. CEUR Workshop Proceedings. 917.