Machine Learning & Qualitative Analysis

Predictive prowess of machine learning harnessed through qualitative analysis .

Alex, Mayuresh, Rucha and Swati

Contents

[Introduction 2](#_Toc532980112)

[Project requirements: 2](#_Toc532980113)

[Data Gathering and Exploration 2](#_Toc532980114)

[Data Retrieval and data manipulation 3](#_Toc532980115)

[Data Analysis and Insights 34](#_Toc532980116)

[Conclusion 34](#_Toc532980118)

# Hypothesis: Given a topic for a book, predict the optimal pricing based on category.

# Introduction

Project 3 is the final team activity and is the cumulative overview of the various topics learnt, including of API data pulls, storing and analyzing data, predictive analysis using machine learning and visualization for the chosen topic.

We worked on finding a solution to predicting optimal price point for a n upcoming book based on category, given the topic. The learnings from one category – books , then can be applied to other categories and help forming a pattern to help the seller determine the best price for their offerings. We want to deep dive into data analysis and gain insights on predictive capability of the data via machine learning.

# Project requirements:

Using Machine learning

In addition we also employed:

* Python Pandas
* Python Matplotlib
* Python Flask
* HTML/CSS/Bootstrap
* JavaScript Plotly
* Javascript D3.js
* Google Cloud SQL
* Tableau

# Data Gathering and Exploration

* Data gathering from Amazon product data (source:http://jmcauley.ucsd.edu/data/amazon/links.html)
* Challenges:

The main challenge was to process and clean 20GB of data in json format to fit in pandas and make it usable for models.

A lot of encoding was required before data was ready to be used in the model.

Most of our data is in string format and models required the data in int or float format.

## **Data Retrieval and data manipulation**

* We reached out to UCSD to gather the updated data on Amazon products.
* To avoid skewed data, we had to shrink the data via normalization process.
* Initial plan was to process data into higher level categories. But it this process turned out to be too manipulative for data with respect to accuracy and utility.
* We experimented with feature set and predictor and finally decided to use feature set as category and price. The output variable was user ratings or acceptance.
* Models included - Random Forest Classifier and Deep Learning

Tableau analysis

* Used the cleaned and merged dataset to find a visual description of our hypothesis.
* The output shows high concentration of volume of books sold below $5 price range and closer to a 4 star rating range.
* This helps in narrowing and understanding the categories of books that have higher potential for success.

# Conclusion

Our sample size is greater than 260,000 ratings and products. Although this is a good representative sample size for understanding the correlation and setting up predictive analysis, un aggregated data might go even further in terms of accuracy of the model.

We have focused on making our model robust, so we can use different datasets to get predictions about optimal pricing and ratings data based on qualitative aspects.