XGBoost - A Competitive Approach for Online Price Prediction

Joshua D. McKenney, Yuqi Jiang, Junyan Shao, Matthew A. Lanham

Purdue University Krannert School of Management

mckenney@purdue.edu; jiang300@purdue.edu; shao52@purdue.edu; lanhamm@purdue.edu

Abstract

This study generates price prediction suggestions for a community-powered shopping application using product text features. Our motivation for this study is it can be a great competitive advantage for companies or individual sellers to have highly accurate pricing decision-support. After doing some EDA, we created text features with above/below average prices to identify the most important features. We used R and Kernels to perform text analysis to generate features from unstructured product features, then used XGBoost and Linear Regression to dynamically predict product price. XGBoost was able to handle over 2,000 brands data while Multiple Linear Regression was not. XGBoost achieved the best performance, with a 0.513 test set RMSLE.

Introduction

- Dynamic pricing is an important decision-support tool that can give sellers precise price suggestions on C2C selling platforms, which can attract more people to use the platform, thus develop the business.
- The trend of vintage and no-name clothing brings the importance of features other than brand recognition and historical pricing when making price prediction.
- Text analysis is becoming more mainstream in business, and a complement to predictive analytics. Correctly building models on text features are the foundation of an accurate prediction in our study.

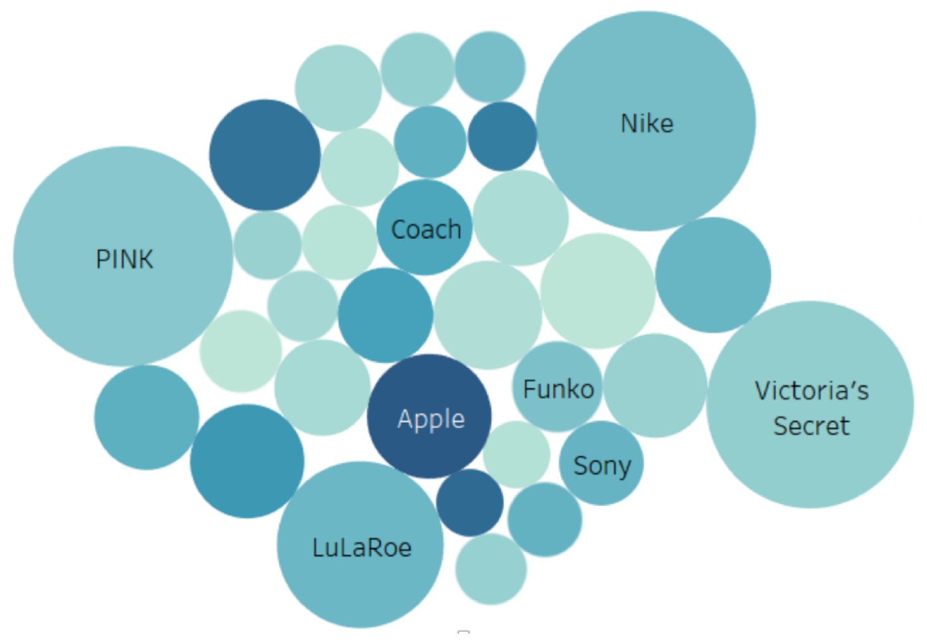

Figure 1. Popular Brand and Prices

Main research question

- How well can we predict the price for an online retailer using textual product features?

Literature Review

We researched on XGBoost and Dynamic Pricing. XGBoost is a popular machine learning methodology frequently used in data analytics and statistics research.

Study	Features
Kannan, Kopalle (2001)	Physical values (ex. Appearance) Non-physical values (ex. Comments of customers)
Bergemann, Välimäki (2006)	Mass market Niche market Social efficiency

Table 1. Literature Review - Features Analysis

Study	Logistic Regression	SVM	Linear Regression	XGBoost	XGLinear	Dart	LOGlm
Li, Yao, Lian, Qiu (n.d.)	✓	✓		✓			
García-Calderón Chávez, Saúl Abraham (2017, July)				✓			✓
Our Study			✓	✓	✓	✓	

Table 2. Literature Review - Methods Used Summary

Methodology

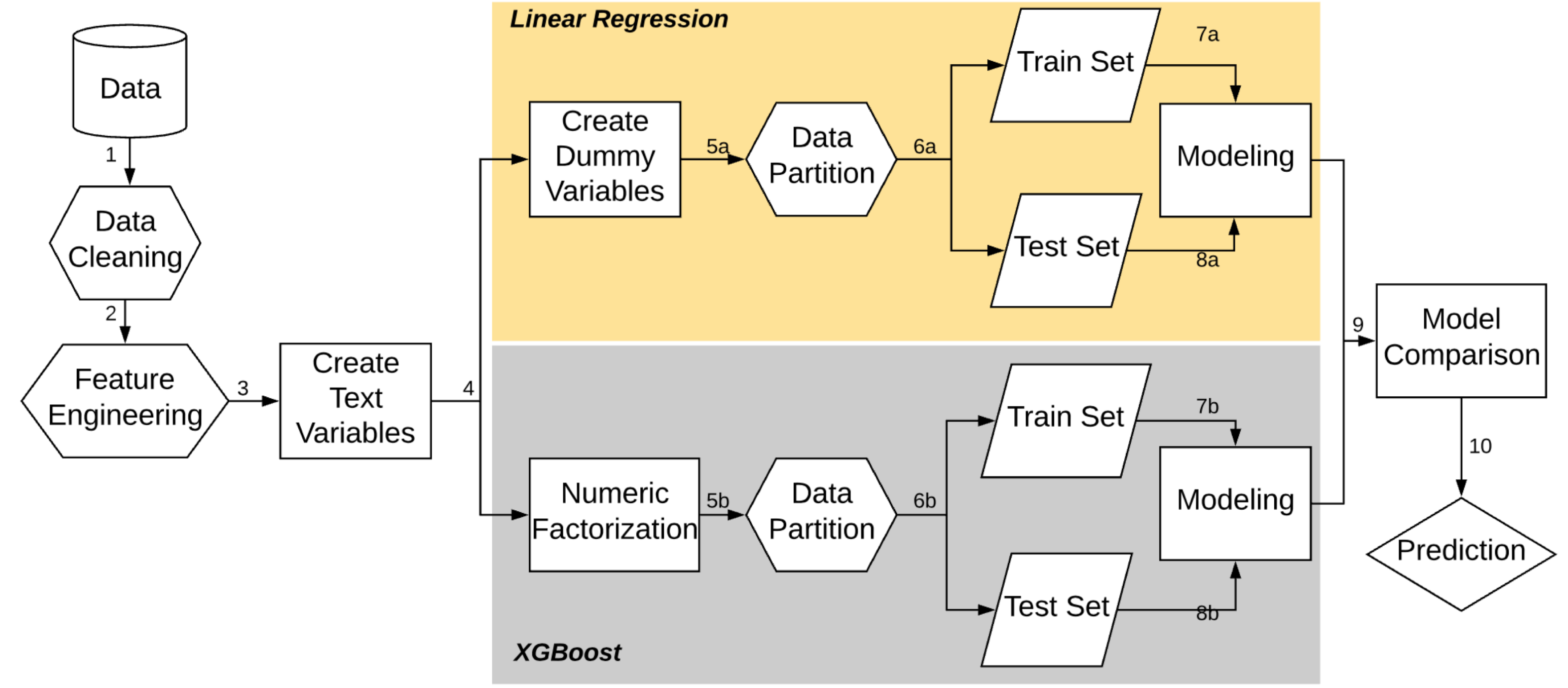

Figure 2. Study Design

Data

- Data from Mercari Kaggle competition, 6 features (exclude ID), 5 features are categorical features → Created dummy variables to do text analysis

Data Cleaning

- Reduced missing value percentage of the dataset from 43% to 0.4%.
  - Missing values all came from “brand_name” column, but most brand names can be found in “name” column.
  - Detected top 10 brand names in the “name” column.

Model Design

Before creating dummy variables, we divided the data set into two parts based on price (higher/lower than average). Text analysis was performed to words/phrases with the highest frequency and separated by number of words combination. Then we partitioned the data into 80-20% train-test sets and used 3-fold cross-validation.

Methodology (Approach) Selection

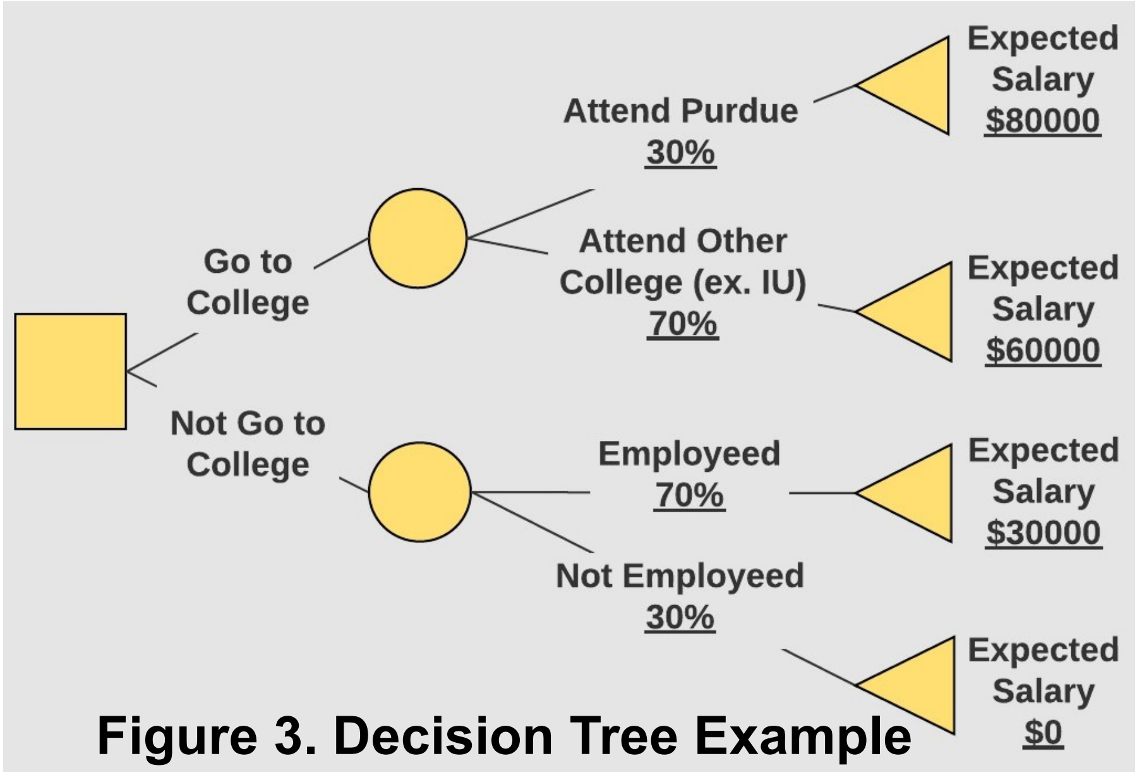

Figure 3. Decision Tree Example

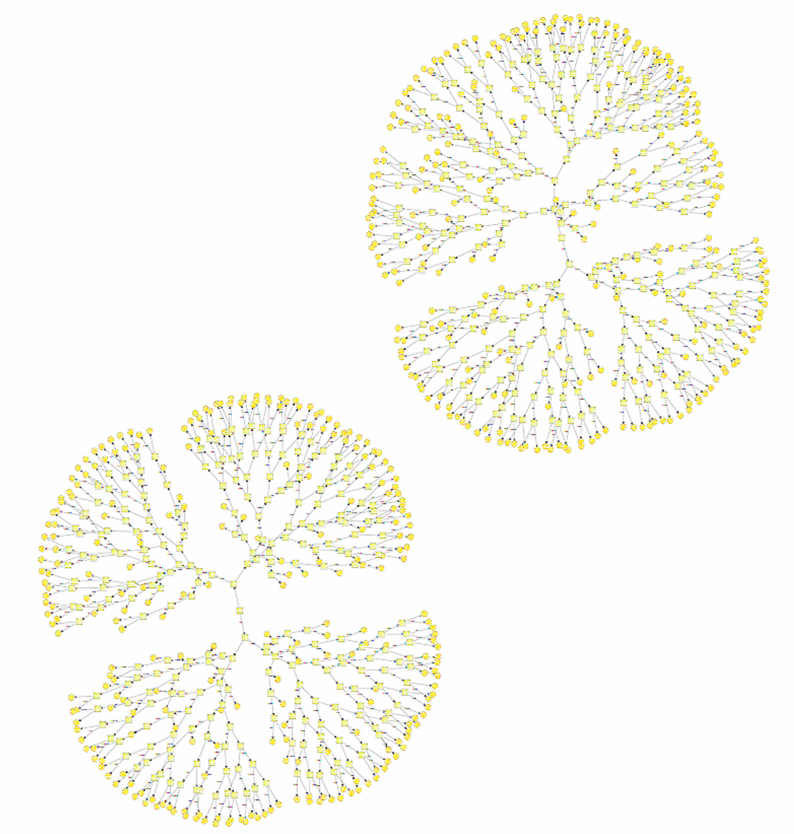

Figure 4. Decision Tree in Our Study

XGBoost is based on the tree boosting system. It provides an efficient way to solve classification and regression problems. The key feature in XGBoost is that it weights the predictors and tries to keep the new decision tree away from the errors made by previous decision trees, thus it strengthens accuracy.

Model Evaluation / Statistical & Business Performance Measures

- Evaluated on overall accuracy, R-Squared and RMSLE
  - RMSLE (root mean squared logarithmic error) is a lower-the-better indicator of model prediction accuracy used in this Kaggle competition.

Results

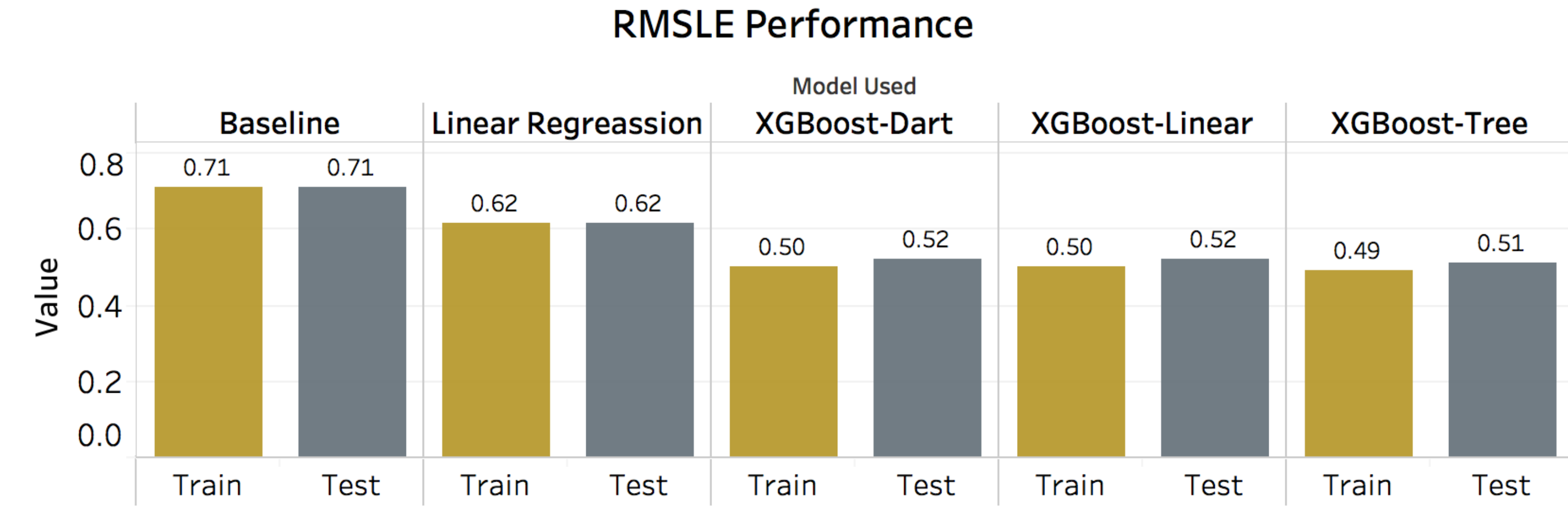

Figure 5. Model Evaluation using RMSLE

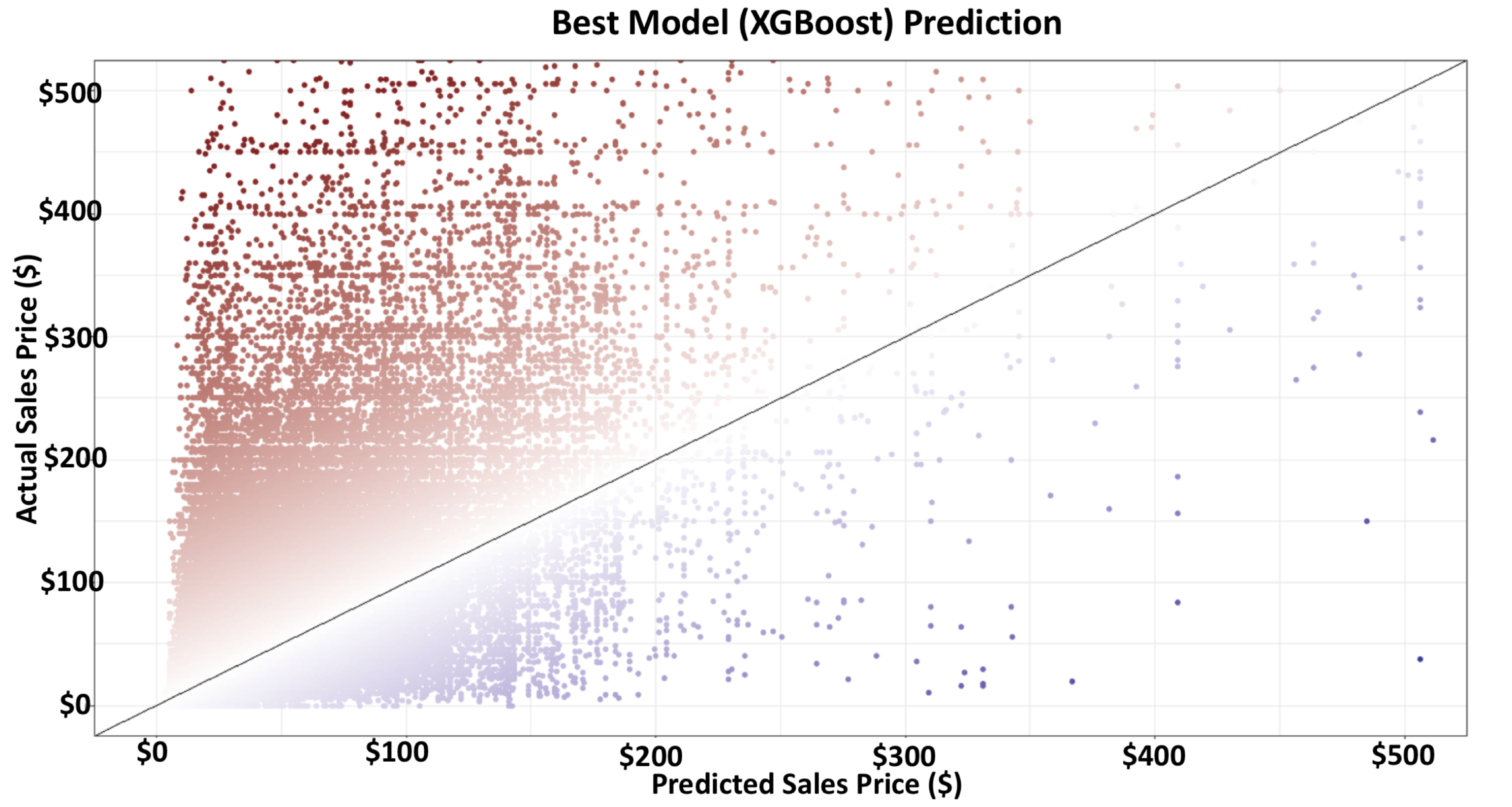

Figure 6. Predicted Price vs Actual Price

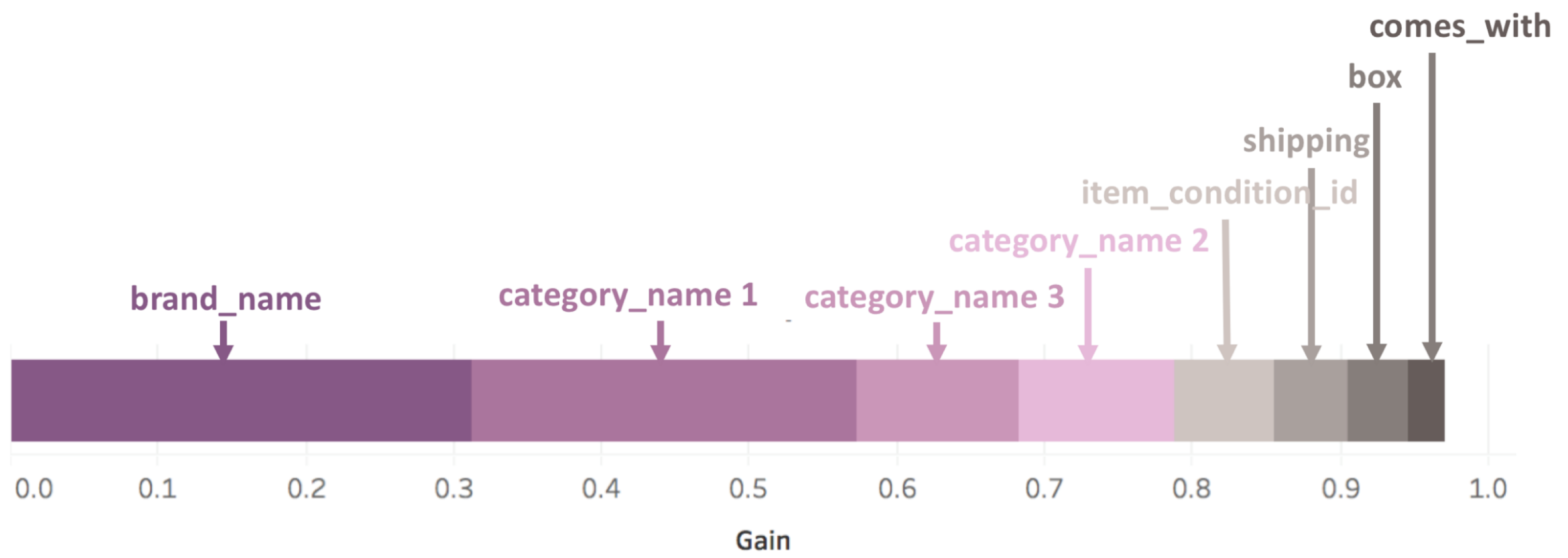

Figure 7. Feature Importance

Conclusions

By precisely and actively predicting the price of a given product based on its various kinds of features, it would be much easier for companies and individual sellers to know about how buyers will value their products.

XGBoost is our best performing model with the best RMSLE. We also proved that text features like item category or product description are very important for accurate price prediction. Business and individual sellers should focus on these features more to save time and make better pricing.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.