XGBoost - A Competitive Approach for Online Price Prediction

Joshua D. McKenney, Yuqi Jiang, Junyan Shao, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
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 features, which is a recent topic of a Kaggle.com competition sponsored by Mercari, Inc. As eBay acquired Canadian data analysis firm Terapeak, the importance of using "big data" and machine learning to improve pricing decision-support in business has been rapidly increasing. By obtaining a solution for price prediction via product features for B2C and C2C online retailers, it will be easier for sellers to sell, and enlarge the selling-shopping community of such user-based marketplaces. It could also be a remarkable competitive advantage for companies or individual sellers having highly accurate pricing decision-support. The authors did some exploratory data analysis, we created text features with above/below average prices for the most important features in the dataset, used R and Kernels to perform text analysis to generate features from unstructured product features, then used XGboost and Multiple Linear Regression to dynamically predict product price. XGBoost was able to handle over 2,000 brands data in our case while Multiple Linear Regression was not able to. XGBoost achieved the best performance, with a 0.513 test set RMSLE.

Keywords: Price Prediction, Product Features, Regression Analysis, Text Analysis, XGBoost

INTRODUCTION

We are seeking solutions for accurately and actively giving price suggestions for B2C and C2C retailing. By using text analysis on different product features, we want to give a precise and dynamic pricing prediction on products which are selling online. This would be extremely important for the C2C platforms. The ability of precisely predicts the price of the product they are going to sell will save time for the sellers thus attract more people to join the community and enlarge the population and sales amount. Competition for apps today is on how many people are actively using the app. So we posit that the success of predicting pricing for online product listings will be a splendid chance for a business like Mercari.

According to Wall Street Journal, eBay acquired Canadian data analysis firm Terapeak in December 2017 (Hanly, 2017). The data analysis company is good at predicting supply, demand, and pricing products. This is an important step in developing demand and price prediction of their online listing features and shows the importance of doing so. eBay is hoping this company could help them more in the data analysis field by providing them capabilities to know more about their sellers, customers, and products.

There is also a change on how people are buying products today. Previously, there were not a lot of people who would consider buying used clothes. But "*The Retail Apocalypse Is Fueled by No-Name Clothes*" now (The Business of Fashion, 2017). Fashion of vintage is bringing the old and no-name clothes back, and people are more open on shopping nowadays. The increasing number of no-name clothes brings up the importance of predicting product prices, because they are all different. Previously, pricing could heavily rely on brand recognition and historical pricing, but the no-name clothes are totally different.

The urgent need of building a model for dynamic price prediction has been raised with no doubt. Jointing statistical researches with business sense and bringing different kinds of models to fit the contemporary business problems, business analytics is the most popular and fully utilized field to address this problem. Our research is utilizing business analytics to find a best fitting model to address online price prediction issue and give the best solution for dynamic pricing to businesses and individual sellers. Big data analysis and text analysis are two most practical and important tools within the business analytics field, and they are the tools which direct us to find the solution.

Though the importance of big data analytics and text analysis is gradually increasing, their power has still been undervalued. Only in the last five to ten years have firms began to invest into finding value from such information. From a decision-makers perspective, it might be difficult to image how one could perform machine learning and dynamic prediction using words as input variables.

However, big data analytics techniques, particularly those focused on generating insights from text are growing fast as is becoming common-place data analysis by non-technical team members. Social media has provided a massive opportunity for data scientists to keep learning on text using more sophisticated techniques. For example, Professor Chen Ying was able to develop a text mining technique to extract and automatically perform text analysis on public data (Phys Org, 2017). Though the amount of data is much less, the use of text analysis is extremely crucial during our prediction of pricing, and thus leads us to our primary research question in this study: how well can we predict the price for an online retailer using textual product features?

We structure our paper by performing a review of the academic literature to frame our research questions, discuss the data we used in our study, outline our methodology, develop our models, and summarize our results.

LITERATURE REVIEW

Our main focus is on how to predict the price for an online retailer. Throughout the literature, we found that our target problem is similar to that posed in dynamic pricing strategy. Dynamic pricing is a term referring to find the optimal price for goods, especially online goods. The goal is to fit the price into the product's features, such as supply and demand, but also brand and quality. The basic idea is to determine the best price by analyzing product characteristics, which is why we believe this research area is similar to what we need to address in our study.

To expand this question, we further research on the possible features that might be used to estimate a price. In the paper "*Dynamic Pricing on the Internet: Importance and Implications for Consumer Behavior,*" the authors indicate that when predicting the price for certain products, not only the physical values need to be considered, such as the appearance of the product, but also the information behind that product, such as the comments of the customers should be considered. In the paper "*Dynamic Pricing of New Experience Goods,*" the authors suggest that whether the market is a mass market or a niche market is really important for the price determination and social efficiency is another factor, since word-of-mouth has a significant power.

To analyze these critical product features, we researched and found that XGBoost is a popular machine learning methodology that has had success in this space and is frequently used in data analytics and statistics research. XGBoost is based on an end-to-end tree boosting system. It offers a faster and more accurate way to solve classification and regression-type 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 its accuracy. This idea of re-weighting predictors where errors occur is the key idea behind all boosting algorithms. While

XGBoost is used in many fields, price prediction by XGBoost has had success. In the paper "*Predicting Buyer Interest for New York Apartment Listings Using XGBoost*" researchers tried several different methods to obtain the best pricing model, including logistic regression, support vector machines (SVM), and XGBoost. In their study, XGBoost provided the best solution.

Table 1. Literature Review Methods Used Summary

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			✓	✓	✓	✓	

During the review of other professional research papers, we found that XGBoost is well adapted for dynamic pricing problems, which is relative to our purpose - predict the price for online retailers. The paper "*Pricing Recommendation by Applying Statistical Modeling Techniques*" successfully uses XGBoost to predict price.

In their paper, the given objective function is:

$$Obj(\theta) = L(\theta) + \Omega(\theta)$$

where L is the training loss function that deals with the extent that the training data can predict the model accurately, and Ω is the regularization term to deal with over-fitting issue of models. The paper indicates that the common way to do training loss function is mean squared error, which is the same objective function that one would use with traditional ordinary least squares regression.

$$L(\theta) = \sum\nolimits_i (y_i - \hat{y}_i)^2$$

The additional regularization term to the objective function is synonymous to other formulations such as ridge regression that adds a penalty to the size of the estimated parameter coefficients (i.e. $\hat{\beta}_j's$) by taking the sumproduct of the squared $\hat{\beta}_j's$ corresponding to the features. In the popular least absolute shrinkage and selection operator (LASSO) case, it takes the sumproduct of the absolute value of the parameter coefficients corresponding to the features. The idea behind these regularization or shrink penalty terms in the objective function helps to obtain more robust models that have better bias-variance tradeoffs than without them.

In conclusion, XGBoost is a highly accurate technique to do model predictions. Thus, in this paper, we focus on using XGBoost as a main methodology to solve our research question.

DATA

We used the data set from Kaggle competition "*Mercari Price Suggestion Challenge*." We created many dummy variables in order to do text analysis on the product features. Table 2 shows a data dictionary of the features provided by Mercari. Please refer to the METHODOLOGY section for the creation of dummy variables.

Table 2: Data used in study

Variable	Type	Description
id	Numeric	Id of the listing
name	Categorical	The name of the product
item_condition_id	Numeric	The condition of the items provided by the seller
category_name	Categorical	Category of the product
brand_name	Categorical	The brand name for the product
price	Numeric	The price for the product in USD
shipping	Categorical	The indicator of shipping paid by seller or buyer
item_description	Categorical	The full description of the product

Exploratory Data Analysis

The price range of the dataset is between $0 to $2009, with the average price being $26.74 and median price $17. Ninty-five percent of the prices are at or below $75, and the mode price is $10. The price variable as showed in Figure 1 is heavily right-skewed, which would influence our prediction. Thus, we transformed the price variable by using the logarithm of prices in order to train our models under the balanced data. After taking the logarithm of prices, the distribution has a more Gaussian distribution as showed in Figure 2.

Figure 1: Price Allocation (Range $0 - $100)

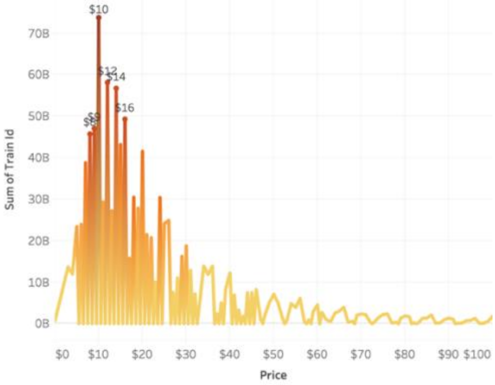
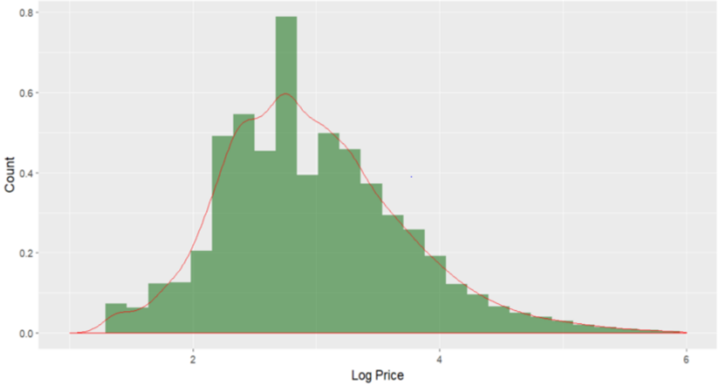


Figure 2: Log Price Allocation



There are 19 categories with greater than 1% of the total products as shown in Figure 3. Within those 19 categories, there are 10 categories under the Women's category. The Women's category accounts for 54% of all records whereas the Men's category accounts for only 8%. There are over 42% of Brand Names that had missing values. Excluding the missing values and the most used Brand Names are PINK, Nike, Victoria's Secret and LulaRoe as shown in Figure 4.

Figure 3: Category Percentage (Greater than 1%)

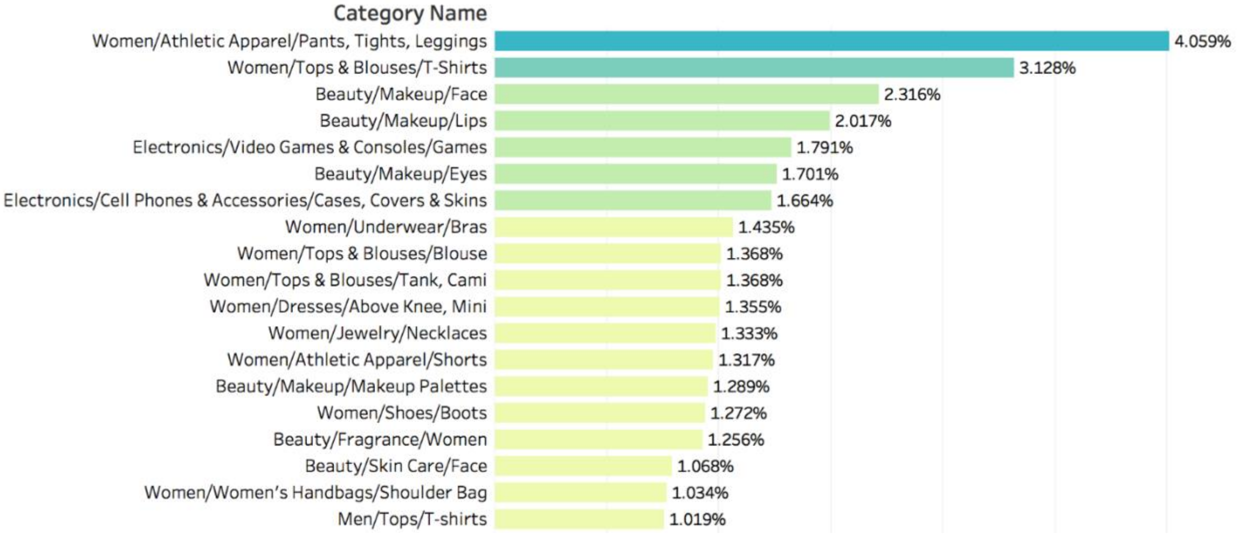
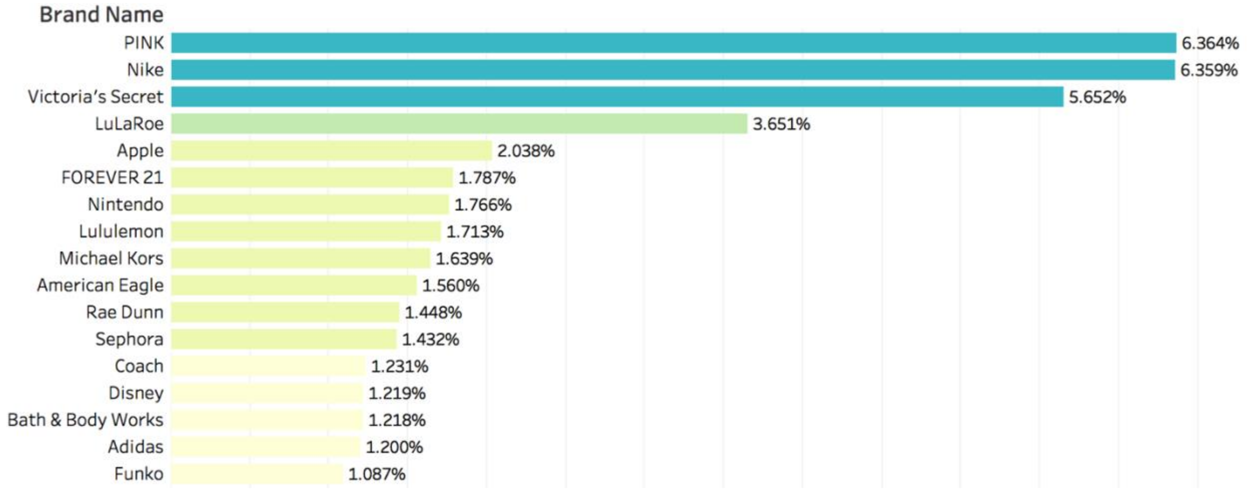
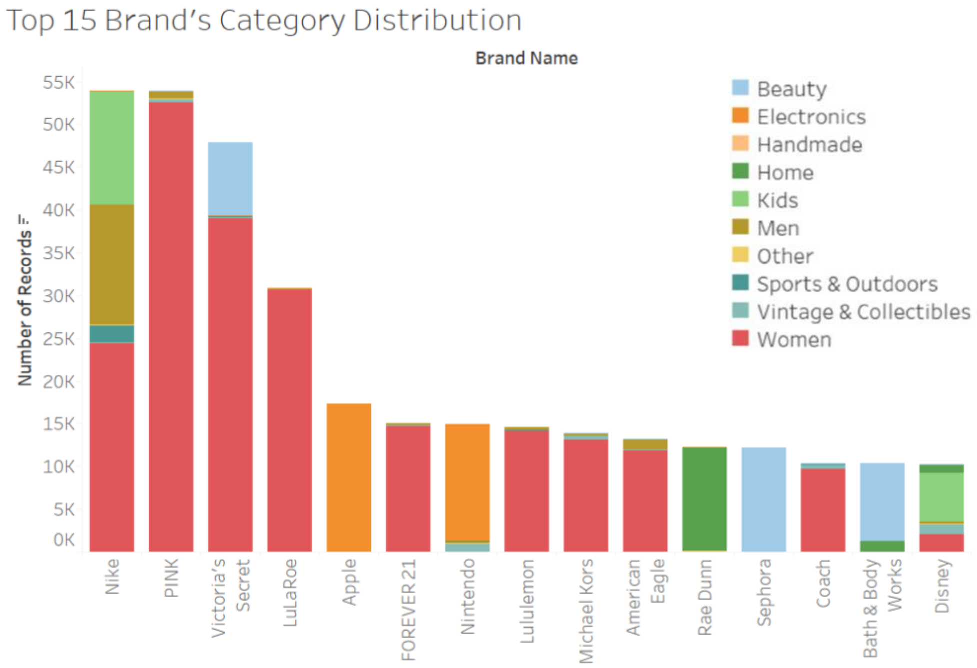


Figure 4: Brand Percentage (Greater than 1%, Exclude Null)



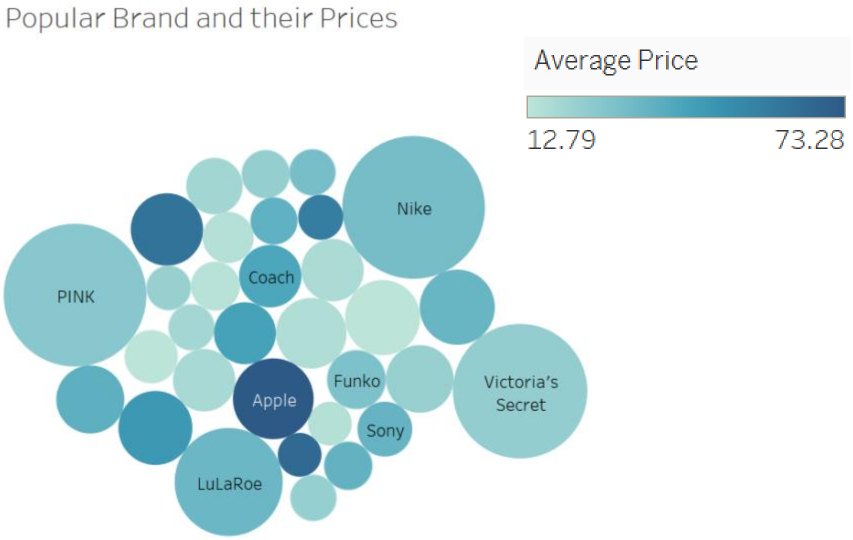
Many of the top selling brands were found mainly in Women's clothing as shown in Figure 5.

Figure 5: Top 15 Brands' Category Distribution



For brands that primarily sell Electronics, such as Apple and Nintendo, tend have above average prices of $73.30 and $34.70 respectively, compare to the average price is $26.74 and the median price is $17 as shown in Figure 6.

Figure 6: Top Brands and their Average Prices

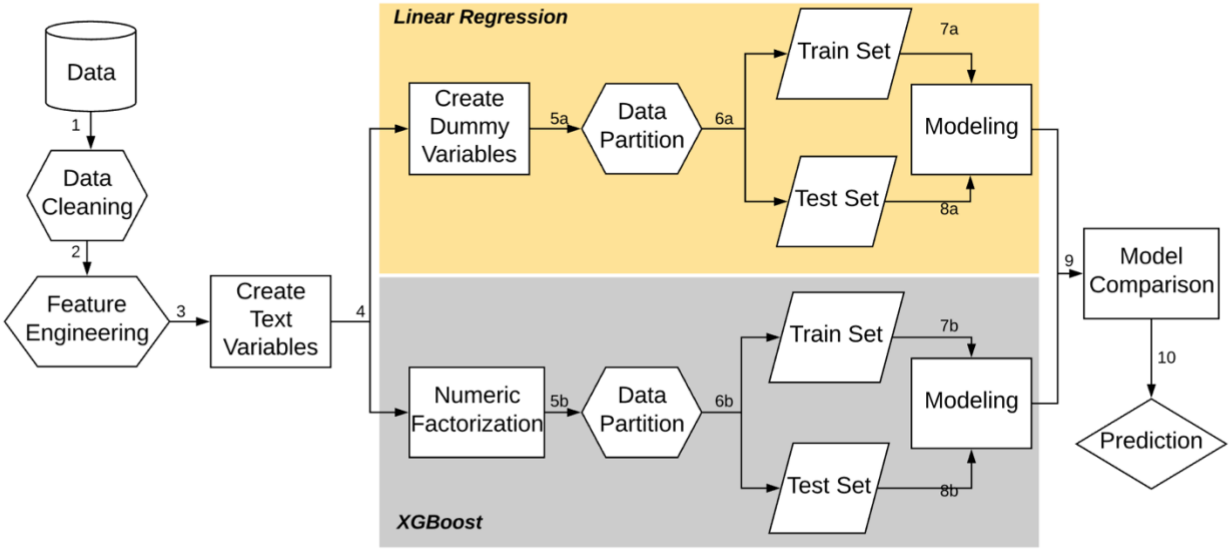


METHODOLOGY

We used multiple regression as our baseline model and variations of XGBoost models as our more sophisticated alternatives. When training these models we used three-fold cross validation on the multiple regression model, and we tried tree, linear, and dart tree boosters for the XGBoost models. We randomly partitioned the original data in an 80-20% train-test sets.

Root Mean Squared Logarithmic Error (RMSLE) is our primary statistical performance measurement. RMSLE is a lower-the-better indicator of model predictive performance specified in the Kaggle competition. We also used the more popular adjusted R-squared as our secondary performance measurer. Adjusted R-squared indicates the percentage of data been explained by the model. Figure 7 outlines the predictive modeling process we applied in our study.

Figure 7: Predictive Model Flow Chart



After performing initial exploratory data analysis, we found that 43% of our data has missing values, and all the missing values come from the "brand_name" column. We also found that for those data with missing value in the "brand_name" column, most of them have the brand name within the "name" column. Knowing this, we detected if one of the top 10 brand names is in the "name" column, and if there were, the observation was given the correct brand name value to the "brand_name" column. By doing so, we cleaned the data and reduced missing values to 0.4%, which significantly helped our model with the hold out set.

To do a regression problem mainly consisting of text, we needed to create lots of dummy variables. To generate dummy variables, we first divided the dataset into two parts based on the prices. One part of the dataset includes the products with a price higher than the average price. The other part includes the products with a price lower than the average price. Text analysis was performed to find those words/phrases with high frequency with one, two, and three-word combinations. These are commonly known as uni-grams, bi-grams, and tri-grams. Comparing the ratios of the frequency of those words showed up in both the above-average-price dataset and the below-average-price dataset. We finalized the text features selection by selecting those words/phrases with high or low ratios as showed in Figures 8 - 11.

Figure 8: Three-word Item (Tri-gram) Description Above Median Pricing

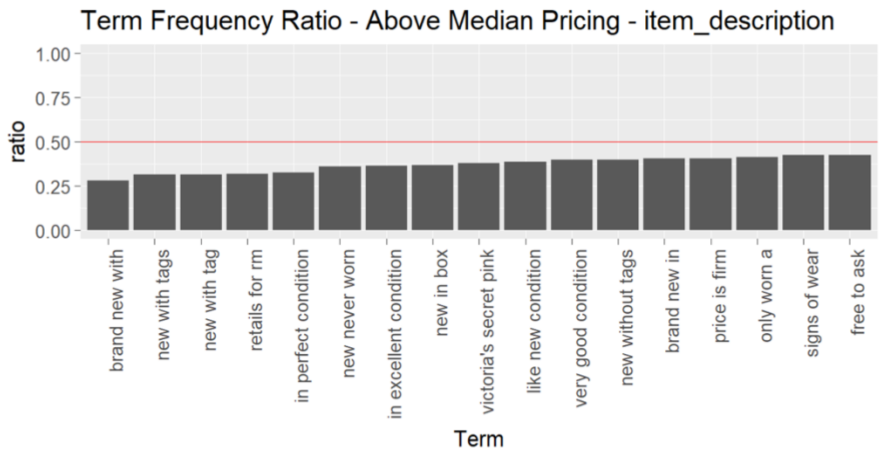


Figure 9: Three-word Item (Tri-gram) Description Below Median Pricing

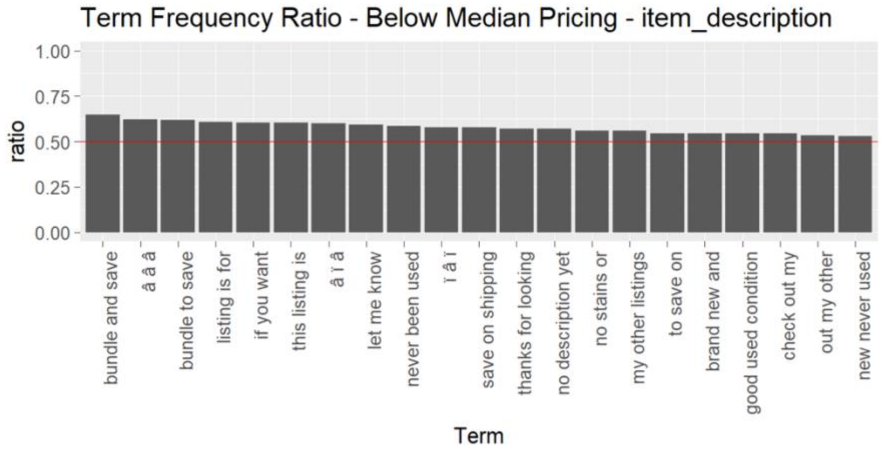


Figure 10: Two-word Item (Bi-gram) Name Above Median Pricing

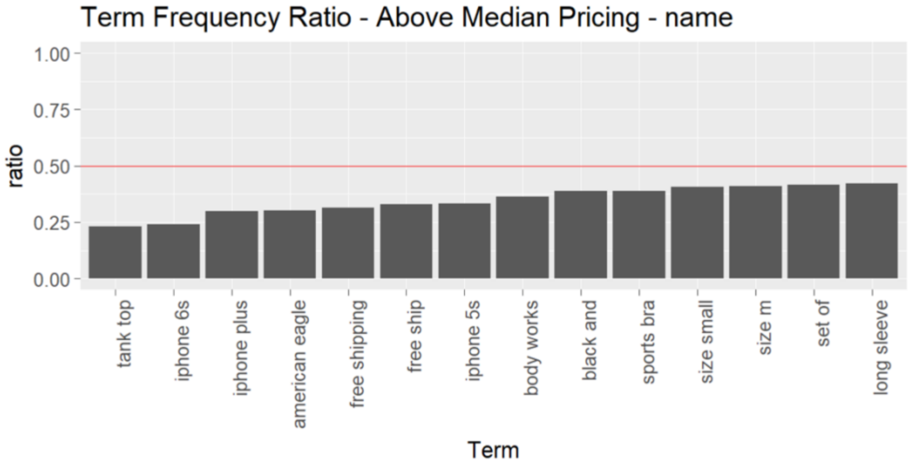


Figure 11: Two-word Item (Bi-gram) Name Below Median Pricing

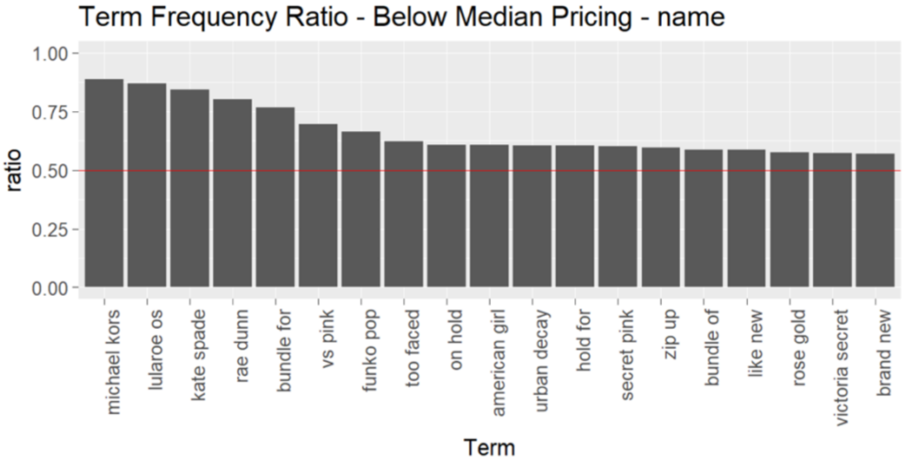
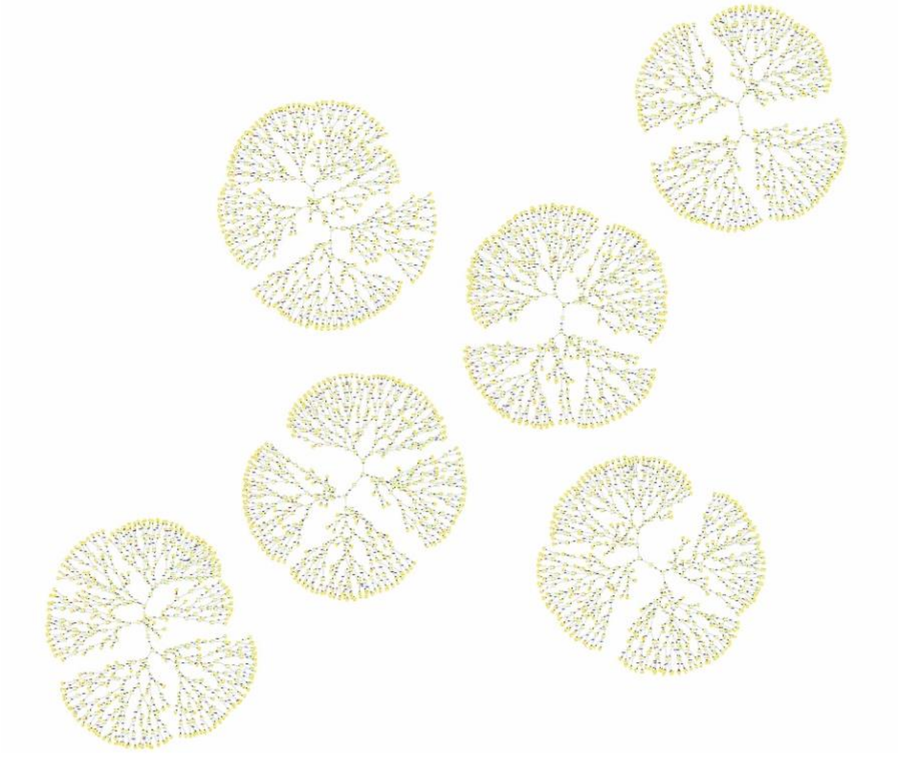


Figure 12: Decision Tree (part) of Our Model



Due to the huge amount of data in our dataset, we generated Figure 12 to show only part of our decision tree and it gives a general idea of our model. It is hard to see each node since there are tons of nodes in this tree. However, this is an overview of how XGBoost is used in our later online price prediction model.

MODELS

Multiple Linear Regression

Multiple linear regression models are models that are a function of several explanatory variables. These parametric models are easy to interpret as each variable's estimated parameter coefficient provides the partial effect that the feature has on the price. The primary drawback of regression is that this model is often not as accurate or versatile as other more complicated predictive algorithms. The reason being that for multiple regression is that a flat plane is fit to the data in p-dimension space, rather than a curved surface that can capture the training data more accurately. This type of model often plagued by bias in the bias-variance tradeoff as it is often too simple of a model when attempting to explain more complicated relationships. The good thing about it being simple is you do not have to worry about overfitting the model to the training data, which would eliminate model generalizability, which is key in the predictive modeling process. Below is the multiple linear regression model we obtained in our study.

Multiple Linear Regression Formula:

$$\begin{aligned} Price = 2.95 &- 0.09(item\ codition\ id2) - 0.15(item\ codition\ id3) \\ &- 0.36(item\ condition\ id4) - 0.28(item\ codition\ id5) - 0.29(shipping) \\ &+ 0.32(cat1Electronics) + \cdots \end{aligned}$$

XGBoost

XGBoost stands for extreme gradient boosting. It is similar to gradient boosting machines, but more efficient and runs much faster than those similar models. Although the model limits input to be only numeric variables and must be in a matrix format, gradient boosting machines are usually relatively accurate when the model is not overfit. The lack of interpretable of these models can be one of the drawbacks that should be considered when developing models that support pricing decisions. The ability of the XGBoost model to handle all 4,809 different Brand Names and all 104 levels of Category 2 and 669 levels o Category 3 variables as numeric factors, compared to Multiple Regression model, greatly helped reduce the RMSLE.

XGBoost Formula:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i)$$

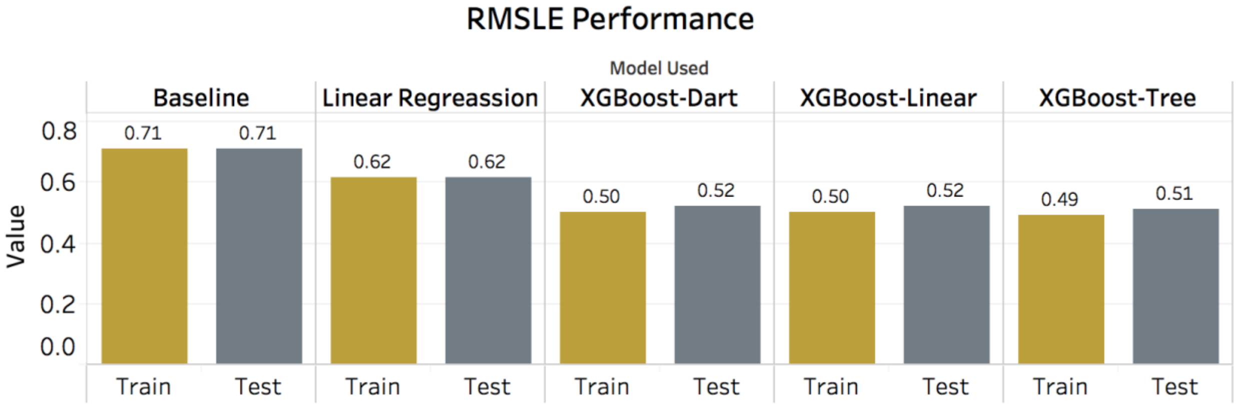
(DMLC, 2016)

RESULTS

RMSLE stands for root mean squared logarithmic error, which a lower-the-better indicator of model prediction accuracy used in this Kaggle competition. We calculated the baseline RMSLE based on the average price in the original dataset, which was 0.7470036.

Figure 13 demonstrates that the neither model overfit to the training data, and XGBoost outperformed linear regression (0.5135 to 0.6169) in RMSLE on the test set.

Figure 13: RMSLE Performance by Models



The XGBoost_tree model performed the best among all models we tried. The result from XGBoost_tree model has the lowest RMSLE and highest R-sqaured as showed in Figure 14. We believe that the XGBoost_tree model will be the best to use for pricing decision support.

Figure 14: R-Sqaured Performance by Models

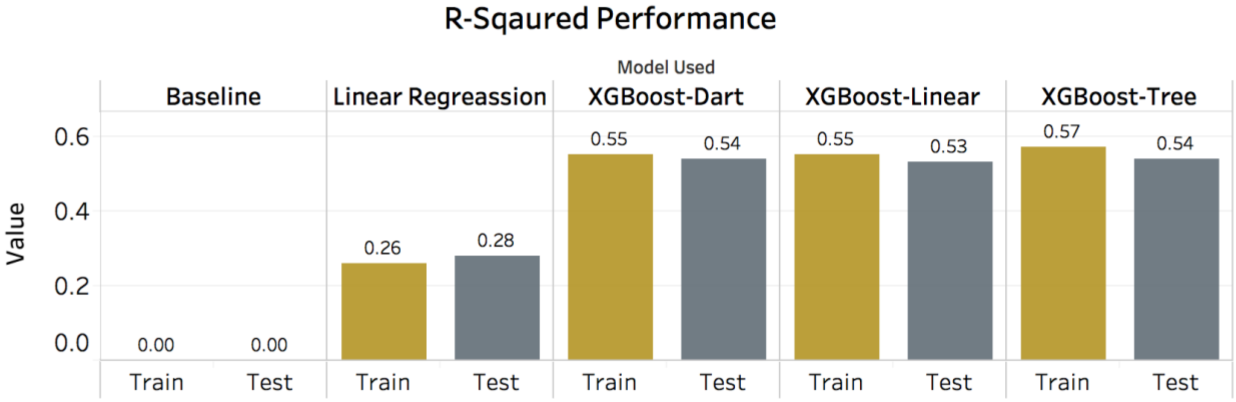
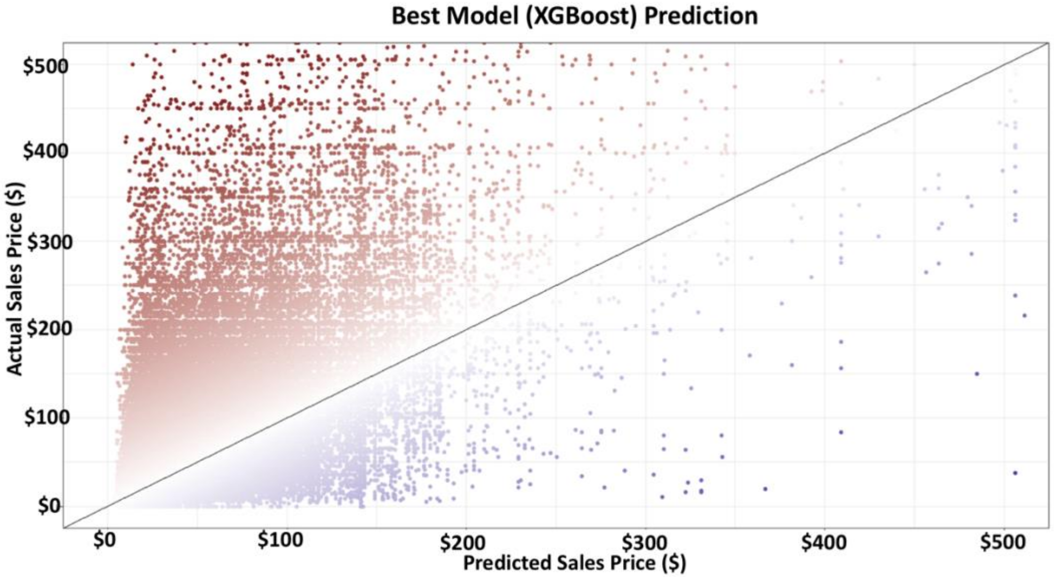


Figure 15 indicates the prediction of our best model XGBoost. Within the price range, lower than $200, the prediction is fairly accurate. But the model tends to over predict than under predict the prices.

Figure 15: Predicted Price vs Actual Price using Best Model (XGBoost)



Although we obtained a relatively satisfied result from the XGBoost model, as previously stated, the results from it are rather hard to interpret. It is our job and it is very important to use those vital functions to help the decision-makers to understand results. The results of the model will be very useful and helpful for any company who wants to find their price drivers are and to predict the resulted price of their current price drivers.

CONCLUSIONS

By precisely and actively predicting the price of a given products based on its various kinds of features, it would be easier for companies and individual sellers to know about how buyers will value their products. For business, they can understand their price drivers more. For individual sellers on a community-based shopping application like Mercari, it would less time-consuming to sell.

We found that transforming text to input features is actually easy to do using popular analytics tools such as R and Kernels. These features can be learned by sophisticated machine learning algorithms such as XGBoost. While popular linear regression did not perform well, XGBoost did. How well can we predict the price for an online retailer using textual product features? Following our process, we were able to predict prices with adjusted R-square of 50%. That means that using textual features explains half of the variation in the price. This provides additional evidence to the previous studies we read that suggests these features can be important. We achieved a 0.513 test set RMSLE compares to the wining team RMSLE of 0.378.

We plan to extend this pricing project to provide better decision-support for pricing decisions. For example, we know we have the ability to predict prices with a reasonable degree of accuracy, but how could these forecasts be strategically used to identify price sensitive products. At the end of our research, we tried LightGBM and got a similar result as XGBoost. But we found that LightGBM was faster and more efficient on making predictions for dynamic pricing with lots of features. We think that it's a direction to do future research since efficiency and time management is crucial when it comes to implementation to business and support business decisions in a timely manner.

REFERENCES

Chachimouchacha. (2017). *BEGINNER'S GUIDE TO MERCARI IN R - [0.50586]*. Retrieved from https://www.kaggle.com/jeremiespagnolo/beginner-s-guide-to-mercari-in-r-0-50677

Dirk Bergemann and Juuso Välimäki, "Dynamic Pricing of New Experience Goods," Journal of Political Economy 114, no. 4 (August 2006): 713-743.

DMLC. (2016). *Introduction to Boosted Trees*. Retrieved from http://xgboost.readthedocs.io/en/latest/model.html

García-Calderón Chávez, Saúl Abraham (2017, July). Pricing recommendation by applying statistical modeling techniques. Retrieved from https://upcommons.upc.edu/handle/2117/109814

Hanly, Ken. (2017, Dec.13. *Internet giant eBay buys Canadian data analysis firm Terapeak*. Retrieved from http://www.digitaljournal.com/business/internet-giant-ebay-buys-canadian-data-analysis-firm-terapeak/article/509906

Mercari. (2017). *Mercari Price Suggestion Challenge - Can you automatically suggest product prices to online sellers?* Retrieved from https://www.kaggle.com/c/mercari-price-suggestion-challenge

P. K. Kannan, Praveen K. Kopalle. (2001). *Dynamic Pricing on the Internet: Importance and Implications for Consumer Behavior*. International Journal of Electronic Commerce, 5:3, 63-83, DOI: 10.1080/10864415.2001.11044211

Phys Org. (2017, Dec. 8). *Unlocking the power of web text data.* Retrieved from https://phys.org/news/2017-12-power-web-text.html

Roozbehani, M., Dahleh, M., & Mitter, S. (2012). Volatility of Power Grids Under Real-Time Pricing. *Power Systems, IEEE Transactions on, 27*(4), 1926-1940.

The Business of Fasion. (2017, Dec. 11). *The Retail Apocalypse Is Fueled by No-Name Clothes*. Retrieved from https://www.businessoffashion.com/articles/news-analysis/the-retail-apocalypse-is-fuelled-by-no-name-clothes

Walters, Troy.(2017). *A Very Extensive Mercari Exploratory Analysis*. Retrieved from https://www.kaggle.com/captcalculator/a-very-extensive-mercari-exploratory-analysis

Yinan Li, Yan Yao, Zhen Lian, Zhihong Qiu (n.d). Predicting Buyer Interest for New York Apartment Listings Using XGBoost. Retrieved from https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a092.pdf