A Competitive Approach for Online Price Prediction

Comparison Among **Regression-Based**, **Tree-Based** and **Ensemble Models**

Yuntong Lin, Fan Lu, Shengye Guo, Matthew A. Lanham
Purdue University Krannert School of Management
lin1053@purdue.edu; lu714@purdue.edu; guo506@purdue.edu; lanhamm@purdue.edu

DAWN OR DOOM 2018 — EMERGING TECHNOLOGY – RISKS AND REWARDS — Two-Day Conference, Nov. 5 & 6

BIAC — Business Information and Analytics Center

Abstract

This study aims at providing precise price prediction decision-support to help vendors decide prices for items they sell online. It can be challenging to know how much a product is worth, and product attributes can lead to significant price differences. This solution provides a cross-validated empirical-based price estimate for a product, rather than a subjective estimate. Using RStudio, we performed text mining to create model features, machine learning methods to predict price, and demonstrate how under-and-over forecasting price impacts key business KPIs. We discuss the impacts of using these models based on both statistical and business performance measures.

Introduction

Having a decision-support tool that could provide vendors precise pricing suggestions on C2C selling platforms would be very advantageous, especially for firms that compete on price. More precise and efficient pricing could help them attract more customers, as well as maximize their margin.

This project explores the **importance of accurate pricing**, which is important for online retailers. When the forecasted price is higher than actual, the customers will go to other websites, which leads to lost sales. If the forecast price is lower than optimal, the company will suffer from lost margin. This project aims at providing optimized prediction and keeping the lost business and margin lost in control.

Assumptions:

- The price offered in the dataset was indeed the optimal price.
- An over forecast by ε % will lead to **lost sales** (will go to a competitor).
- An under forecast leads to **lost margin**.

Research Questions:

- How can one use textual product features to accurately price a product?
- How do **machine learning models** (Regression-Based and Tree-Based models; LASSO, Random Forest, XGBoost and Ensemble models) perform at predicting price when using textual-derived features?
- How does the business performance (e.g. margin, lost sales) change with changes in model accuracy?

Literature Review

We researched studies on 1) Price Prediction, 2) Text Mining, and 3) Machine Learning models, including regression-based and tree based models.

Linear Regression	**Decision Tree**	LASSO	Random Forest	XGBoost	Neural Network	SVM	Ensemble Model
-Numerical data with lots of features	- classification - Medical diagnosis - Credit risk analysis	-regression based -high dimension or overfitting	-Tree based -High dimension -Feature importance	-Tree based -High dimension -parallel computation	- Images - Video - "Human-intelligence" type tasks like driving or flying - Robotics	- Classifying proteins - Text classification - Image classification - Handwriting recognition	-Better prediction -More stable mode
X	X	X	X	X			X

Table 1. Literature review summary by method used

The novelty of this study lies in how we 1) compared regression- and tree-based machine learning models using textual features for price prediction; and 2) demonstrate the effects using each model would have on business KPIs. Too often this crucial step is missed in connecting the models to the business.

Methodology

Figure 1 outlines our study design, starting from data cleaning, data pre-processing, feature creation and selection, model selection, cross-validation design, and model assessment measures.

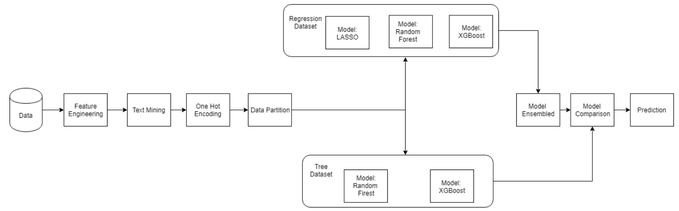


Figure 1. Study Design

■ **Data** (Source: Mercari Kaggle competition)
Target feature: Price; categorical features included: shipping and condition
Textual features: name, category name, brand name, description

■ **EDA, Data Cleaning & Pre-Processing**

➢ **Regression Based Models**

1.Using N-gram (uni-gram, bi-gram, tri-gram) for 'Item Description'; Uni-gram for 'Brand Name', select top 20 most used grams and create dummy variables.

2.Detect size, color, usage condition, material, etc. information in 'Item Description' and create dummies from this extra description information.

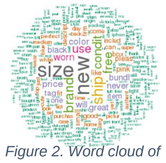


Figure 2. Word cloud of 'Item Description'

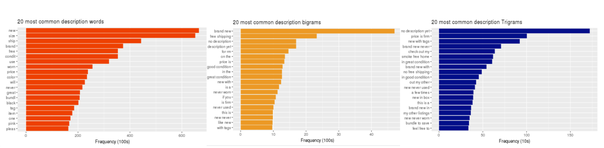
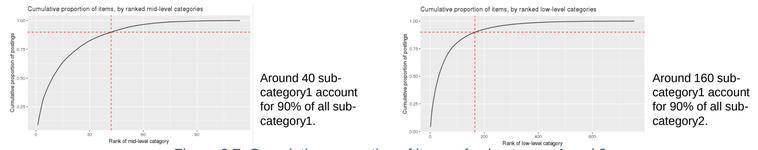


Figure 3,4,5 Top 20 most used grams using Uni-, bi-, tri-gram

3.Split 'Category' into 'general category', 'subcategory1' and 'subcategory2'. created 9, 12, and 7 dummies for each respectively.



Around 40 sub-category1 account for 90% of all sub-category1.

Around 160 sub-category1 account for 90% of all sub-category2.

Figure 6,7. Cumulative proportion of items of subcategory 1 and 2.

4.Create dummies for 'Condition' (1 to 5).

➢ **Tree Based Models**

1.Changed textual features (eg: name, brand, general category, subcategory1, subcategory2) into numeric features.

2.Merge with processed 'Item Description'.

■ **Model Design**

Data was partitioned into a 70-30% train-test set, and 5-fold cross-validation was used to evaluate and compare models, and reduce overfitting.

■ **Methodology (Approach) Selection**

1. Regression Based: LASSO, Random Forest, XGBoost, Ensemble Model.
2. Tree Based: Random Forest, XGBoost, Ensemble Model.

Results: Model Evaluation and Business Insight

■ **Model Evaluation Measures**

The predictive models were evaluated on RMSLE, which represents the test set error of each model.

Reg-Based Models	Set	RMSLE value
LASSO	Train	0.6797616
	Test	0.6856457
Random Forest	Train	0.6292667
	Test	0.6659648
XGBoost	Train	0.6582971
	Test	0.6683317
Ensemble	Train	0.6407201
	Test √	**0.6630945**

Table 2. Comparison of Regression-Based Models

Tree-Based Models	Set	RMSLE value
Random Forest	Train	0.1468371
	Test	0.1583948
XGBoost	Train	0.1575847
	Test	0.1591974
Ensemble	Train	0.1502433
	Test √	**0.1569239**

Table 3. Comparison of Tree-Based Models

After changing to tree-based models, RMSLE drops greatly from 0.66 to 0.15, which indicates that tree-based models perform better in this situation. This low RMSLE of tree-based models is due to the smaller-size dataset. RMSLE will get greater when the size grows, and we estimate that it will increase to around 0.5. We observed, tree-based models are time-consuming and highly demanding on our systems memory.

After comparison among several models, we found ensemble model always has better performance (better prediction & more stable) than others.

■ **Business Evaluation**

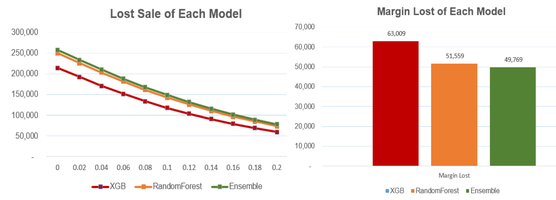


Figure 8, 9. Lost Business and Margin Lost of Each Model

Lost Sale (Lost Business)			
Epsilon (OverPrice)	XGB	RandomForest	Ensemble
0	214,256	250,035	257,865
0.02	192,581	226,064	233,712
0.04	170,820	203,386	210,396
0.06	151,594	181,631	188,403
0.08	133,708	161,454	167,839
0.1	117,960	142,999	149,044
0.12	104,073	125,965	131,770
0.14	90,922	110,674	115,722
0.16	79,676	97,093	101,706
0.18	69,472	84,970	89,326
0.2	60,035	73,995	77,946
Margin Lost			
	XGB	RandomForest	Ensemble
Lost	63,009	51,559	49,769

Table 4. Lost Business and Margin Lost

The ensemble model led to the lowest margin lost (tends to under forecast the least), and XGB tended to lead to the best possible model to reduce lost sales.

Conclusions

This study aims at seeking optimized solutions for efficiently and precisely prediction price for online retailers, as motivated by the Mercari competition.

Business Recommendation

- Our model could provide precise pricing predictions, which not only reduces the risk of pricing higher than actual and leading to lost business, but also reduces the risk of pricing lower and leading to lost margins.
- While the ensemble model could keep margin lost lower than $50,000 and has the best test set RMSLE (0.1569), the firm might be more concerned with reducing lost sales, thus the XGB should be used. The difference in statistical performance compared to the ensemble model is insignificant (0.1591 vs. 0.1569).

■ **Important Findings in Methodology**

- When a categorical feature has too many levels, we can transfer the categorical values into nominal features and use tree-based models. Tree-based models could greatly reduce the error of prediction.

Future Studies

- Latent Dirichlet Analysis to improve analysis on 'Item Description';
- Include more keywords into models to improve the accuracy.

Decision Support Tool Prototype

Link to a video demonstration of an R-Shiny prototype for this solution.