E-Commerce Sales Prediction Using Listing Keywords

Stephanie Chen (asksteph@stanford.edu)

1 Introduction

Small online retailers usually set themselves apart from brick and mortar stores, traditional brand names, and giant online retailers by offering goods at exceptional value. In addition to price, they compete for shoppers' attention via descriptive listing titles, whose effectiveness as search keywords can help drive sales. In this study, machine learning techniques will be applied to online retail data to measure the link between keywords and sales volumes.

2 System Design

General structure and process flow of the system are illustrated below.

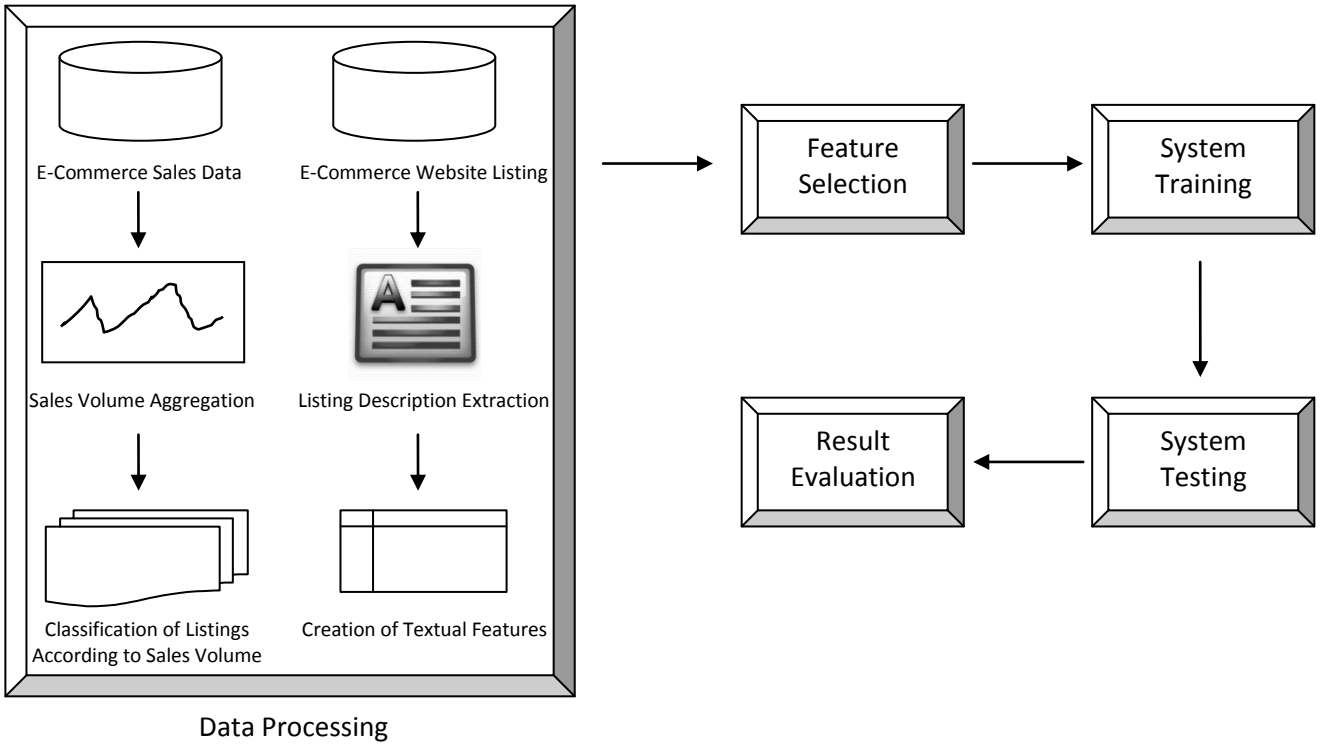

Figure 1: Diagram of System Architecture

2.1 Data Description

The dataset used in the study is a data sample of 500 dress sale listings on the Ali Express online platform, provided by the UCI Machine Learning Repository (Bache and Lichman 2013). It contains sales numbers and 13 product attributes for the merchandise. The sales numbers are snapshots taken from the listing in fall 2013 and represent the sold volumes in the 6 months prior to the snapshot date. The product attributes include style, price, shape, and design categories. Additionally, listing titles and descriptions are collected using a web crawler tool on the URLs provided. The texts are appended to the corresponding dress item.

For this study, a binary label for the dataset is created, based on the average sales volume. 1 denotes a popular item with high-volume sales, and 0 denotes otherwise. Selecting a threshold of 200 for the average sales volume yields a 45/55 split between the hot-or-not sellers, i.e. an item that is sold on average at least once a day in the last 6 months is considered to be a high-sales item for a small online retailer. The modeling experiments for this binary classification problem discussed in the later section are meant to assess the effect of keywords on predicting sales performance.

2.2 Textual Feature Creation

Words are parsed from the texts downloaded from the dress webpages. The uses of the words are tallied up and the most common words (e.g. free shipping, fashion, 2013) minus the stop words (e.g. the, a, and) constitute the bag of 180 words in this study (Ahn and Spangler 2014). Unlike literature and news articles, sale listing title and descriptions are short and succinct to capture shoppers' attention span. How many times a word is used in a listing is not as relevant in this domain as what words are used. Thus, a textual feature is created for each word as a binary variable denoting the presence of the word in the listing.

2.3 Feature Selection

Correlation analysis is performed on the dataset to see which words are highly correlated with the high-sales item listings and how they correlate with each other (Yu et al. 2012). To avoid collinearity, words that convey redundant information are removed. The remaining sales-correlated words become candidate predictor variables for a model scenario later.

Aikaike information criterion (AIC) is also used to perform feature selection. It provides a systematic way of selecting features that minimizes information loss by balancing the tradeoff between goodness of fit and model complexity (Kaya and Karsligil 2010). It serves as a regularization technique that penalizes models with poor fit and models with too many parameters. Feature sets determined from forward search, backward search, and bidirectional search using AIC respectively are examined for statistical significance using p-value. Statistically

insignificant features are removed from each set, and the resulting variables are experimented as separate model scenarios.

Selection Method	Selected Textual Features
Correlation Analysis	women, 2012, mini, cotton, neck, leopard, summer, crew
Forward Search	2012, zipper, solid, cheap, sale, wholesale, selling, piece
Backward Search	slim, color, 2013, short, sale, wholesale, winter, printing
Bidirectional Search	zipper, solid, cheap, sale, wholesale, selling, piece, cute

Figure 2: List of Selected Textual Features

2.4 Model Experiments

There are 6 scenarios to experiment using original data with: I) no textual features, II) all textual features, III) correlated textual features, IV) textual features from forward search, V) textual features from backward search, and VI) textual features from bidirectional search. The first scenario serves as a baseline for comparison with additional textual information on sales prediction. Data is pre-processed for the 10-fold cross validation to be conducted during the model training and testing process. The dataset is randomly split, such that each run uses 90% for training and 10% for testing. Logistic regression and support vector machine are standard models used for binary classification problems and are explored for this study (Culotta 2012). The multiple-model approach provides additional validation of the effect of keywords on sales prediction.

The logistic regression model fit outputs predicted probabilities of the samples having positive labels, and probability of 50% or more is considered a prediction of the item as being popular and having high sales. The SVM model classifies samples according to where they fall with respect to the support vector decision boundary. Initial tests show that SVM models trained with the common Gaussian kernel are overfitting the data and do not generalize well. After experimenting with different kernels, the linear kernel delivers the most reasonably close train and test errors and is thus chosen for the SVM model. It suggests that a linear decision boundary is appropriate for the set of e-commerce sales data used in this study.

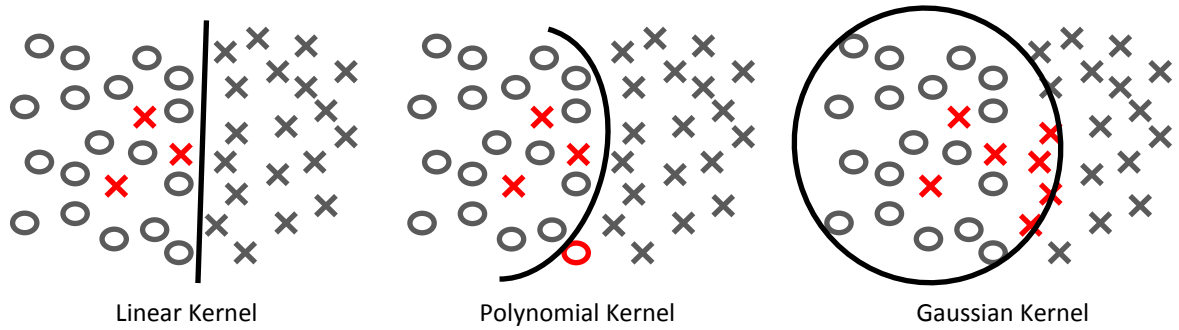

Figure 3: Hypothetical 2D-Illustration of SVM Decision Boundary

Prediction accuracy and sales prediction precision are the metrics used to evaluate the effectiveness of the features and models in identifying the high-sales item listings. Prediction accuracy is calculated as the percentage of predictions being correct out of all predictions made. Sales prediction precision is calculated as the percentage of high-sales (i.e. positive) predictions being correct out of all high-sales (i.e. positive) predictions made. Prediction accuracy and precision of the holdout sets are respectively averaged from the 10 cross-validation runs for each model type in each scenario.

3 Result Evaluation

The first scenario sets up the baseline metrics for other scenarios with textual features to measure up against. The second scenario demonstrates an overfitting situation, in which all textual features are being thrown into the model. Including irrelevant features (i.e. nonessential words) into the model actually deteriorates the model's ability to detect high-sales listings from those that are not.

	Logistic Regression		Support Vector Machine	
Scenario	Accuracy	Precision	Accuracy	Precision
I: No Text	0.6860	0.6449	0.6700	0.6221
II: All Text	0.6060	0.5702	0.6160	0.5677
III: Correlated Text	0.7140	0.6873	0.7120	0.6906
IV: Forward Search Text	0.7200	0.6732	0.6940	0.6318
V: Backward Search Text	0.6840	0.6447	0.6800	0.6301
VI: Bidirectional Search Text	0.7220	0.6880	0.6860	0.6239

Figure 4: Table of Prediction Results by Scenario

The models in scenarios III through VI contain heuristically selected textual features and have been found to outperform the baseline scenario in almost all result categories. The performance lifts in prediction accuracy and sales prediction precision over random behavior and baseline scenario reinforce the notion that certain keywords can be effective indicators of popular online listings with high sales volumes.

4 Conclusion

In this study, a set of online retail sales data is examined to assess the impact of listing keywords on predicting sales performance. Textual features are created from the e-commerce dataset, and candidate predictor variables are selected based on correlation and AIC analyses.

Scenarios are constructed to measure the lift in predictive power of including appropriate textual features into models to classify sales popularity. Experiment results suggest that incorporating certain keywords boosts classifier accuracy and sales prediction precision. Additional A/B testing can be carried out to further investigate whether keywords actually drive higher sales by setting up variant listings with identical content except the exclusion of keywords in the control group.

Other variants to this study can be explored in the future. Instead of employing greedy heuristics in feature selection, models can be optimized by adding regularization directly to the objective functions. Different classifier models like naive Bayes and decision tree ensemble methods that do make many assumptions on the underlying data distribution can also be investigated. Furthermore, the existing system design can be tested on more sales datasets to refine its methodology and prove its validity.

References

Ahn, H. and Spangler, W. S. (2014). Sales Prediction with Social Media Analysis. San Jose, CA: IBM Research - Almaden.

Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Culotta, A. (2012). Lightweight Methods to Estimate Influenza Rates and Alcohol Sales Volume from Twitter Messages. Hammond, LA: Southeastern Louisiana University, Department of Computer Science & Industrial Technology.

Kaya, M. I. Y. and Karsligil, M. E. (2010). Stock Price Prediction Using Financial News Articles. Istanbul, Turkey: Yildiz Technical University, Department of Computer Engineering.

Yu, X., Liu, Y., Huang, J. X., and An, A. (2012). Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain. Toronto, Canada: York University, School of Information Technology.