Predicting the Conversion Rate of eBay Listings

Amin Ghafouri

1 Imbalanced Data

eBay dataset is largely imbalanced. There are different approaches to address the sample imbalance problem:

1. Balancing the training set through oversampling the minority class, undersampling the majority class, or synthesizing new data for minority class.
2. At the algorithm level through adjusting the class weight (i.e., misclassification cost), adjusting the decision threshold, or making the algorithm more sensitive to minority class.

If we over-sample the minority, we risk overfitting. If we under-sample the majority, we risk missing aspects of the majority class. An alternative solution is to use Synthetic Minority Over-sampling Technique (SMOTE). In particular, recent work shows that balancing the dataset with SMOTE and training a gradient boosting algorithm on the balanced set provides a highly accurate predictive model. However, the main issue with this approach is that it is not applicable to the case of predicting probabilities, as in our case.

At the algorithm level, a reasonable approach is to increase the weight of minority class to $\frac{\text{number of 1 instances}}{\text{sum of 0 instances}}$. Another solution is to use algorithms that are more sensitive to minority class such as gradient boosting algorithms. In particular, GBRT and XGBoost are shown to perform well in the presence of unbalanced classes (e.g., as observed in Santander Kaggle competition). In this project, we use a state-of-the-art XGBoost classifier which provides many advantages over other approaches (e.g., scalable, able to handle heterogeneous data, and high predictive power).

2 Preprocessing

Item Name & Description: We assume names and descriptions are preprocessed, i.e., stemmed, lemmatized, removal of stop words and typos, etc. We can use Bag of Words (BOW) or term frequencyinverse document frequency (tf-idf) to represent item name and description. In a BOW approach, each element represents the frequency of a specific word. This frequency can be either boolean, indicating whether the word appears in the text or not, or numerical, indicating the number of occurrences in the text. The distance between the

BOW vectors can be used to find similar items. In tf-idf, a statistical measure is used to evaluate how important a word is to a document in a collection. In this project, we use tf-idf representation as it is more powerful and can achieve better results.

Features: Our prediction model is based upon the features item name, item description, number of clicks per sale, seller ID, item price, listing attributes

We could also use item's fair market price. Note that to estimate item's fair market price, one can use similar items sold as it reveal important demand measurements. This can be highly beneficial since as stated in [1], the ratio of initial total price to the median total price of recently sold similar items is highly correlated with conversion rate. Seller ID is also considered as it is a strong indicator of conversion rate.

3 Learning

We use XGBoost, which is an optimized distributed gradient boosting method designed to be highly efficient, flexible and portable. Gradient boosting methods are know to work well for non-linear heterogeneous classification/regression models with large number of factors. The result is stable even when some important factors are missing at runtime. Note that other approaches can also be used. For example, weighted random forest is also shown to perform well, especially if the out-of-bag estimate of the accuracy from random forest is used to select weights [2].

4 Evaluation

We use 100 decision trees to train our predictive model. We select a small value for learning rate as small values of learning rate will help prevent the algorithm to over-fit the training data. Further, validation sets are sampled without overlapping with the training data.

Since the sample is imbalanced, the overall accuracy is not an appropriate measure of performance. This is because a trivial classifier that predicts every case as the majority class can still achieve very high accuracy. Therefore, we need use other performance metrics such as precision and recall, false positive rate and false negative rate, F-measure and weighted accuracy. In this report, we use the area under the ROC curve between the predicted probability and the observed target.

Given the evaluation scenario described above, and by splitting the dataset with a 3:1 ratio, our method obtains an AUROC of 0.902. Moreover, for each row in the validation data, our code provides the conversion probability of the corresponding item.

5 Improvements & Extensions

- Using an ensemble of XGBoost's as used by the top algorithm in Santander competition [3].
- Designing a predictor to estimate the item's fair price, and using the estimated value as an input to the conversion rate predictor
- Employing other powerful learning algorithms such as deep neural networks and weighted random forests.
- Performance and runtime improvement using feature selection and dimension reduction techniques.

References

[1] Ted Tao Yuan, Zhaohui Chen, and Mike Mathieson. Predicting ebay listing conversion. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1335–1336. ACM, 2011.

[2] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110, 2004.

[3] Haoyue Liu and MengChu Zhou. Decision tree rule-based feature selection for large-scale imbalanced data. In *Wireless and Optical Communication Conference (WOCC), 2017 26th*, pages 1–6. IEEE, 2017.

[4] Liyan Chen and Wen Liang. Airbnb price prediction using gradient boosting. 2016.

[5] Dennis Van Heijst, Rob Potharst, and Michiel van Wezel. A support system for predicting ebay end prices. *Decision Support Systems*, 44(4):970–982, 2008.

[6] Shubharthi Dey, Yash Kumar, Snehanshu Saha, and Suryoday Basak. Forecasting to classification: Predicting the direction of stock market price using xtreme gradient boosting.