BOĞAZIÇI UNIVERSITY

STATISTICAL LEARNING FOR DATA MINING
IE 582

Gender Prediction in E-Commerce

Authors:
Kürşat GÜRSOY
Y. Harun KIVRIL
Kadir PEHLİVAN

24 January 2022

Introduction

Understanding the customer behavior and customizing its marketplace accordingly is one of the main challenges that online retailers encounter. It is possible for an online retailer to increase its revenue by customizing its marketing strategies to attract customers from different segments. Each customer has its own preferences about concepts such as security, loyalty and price sensitivity. These preferences may significantly differ between genders[8]. Therefore, a data holding the gender information of each customer is very valuable. However, in most cases this information is either not available or it is not reliable due to privacy and security concerns of customers. Therefore, there is a necessity of identifying gender information accurately. After this aim is reached, gender information can be extended and it can be efficiently used for marketing and related purposes. In this project, missing gender information of an online retailer users, whose activity data is available, is aimed to be predicted with a model created by using the activity data of users that have filled gender information. For this purpose, new features are created to enrich the given transactional train data which has 2.955.345 instances and 19 features in total. During the feature engineering steps, both train and test data sets are rearranged in a way that they contain informations for each customer in each instance. At the end of data preparation, the train and test sets that have 5618 and 2380 instances with 131 features, respectively, are obtained. Later, different classification models were learned on the manipulated train set and they have been compared on the test set with respect to the metrics mentioned in the project statement, AUC(Area under the ROC curve) and Balanced Accuracy. In final, the model with the best performance is appeared to be the one trained with Light Gradient Boosting Machine Classifier.

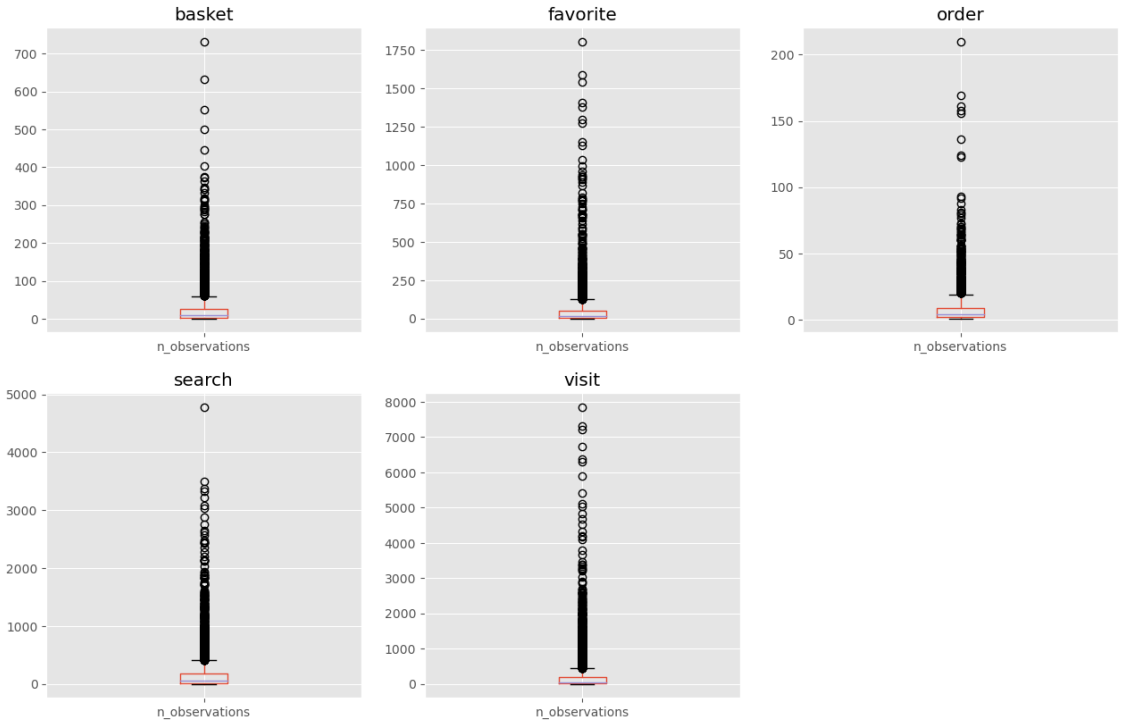

Figure 1: User action observation count distributions

1 Related Literature

In one research, named Customer Gender Prediction System on Hierarchical E-commerce data, a machine learning algorithm of Support Vector Machines(SVMs) classifier for predicting gender information with transaction data is built [5]. Additionally, it is shown that different feature augmentation ways and different ways of feature weighting improve the performance of the machine learning algorithm. The feature augmentation approaches include unique IDs decomposition, context window-based history generation, and extracting identical hierarchy. In this research, each session created from transaction data is modeled as feature vectors, obtained from the session product IDs. Sessions are labeled as male or female and the research has demonstrated that accurate gender prediction has been obtained thanks to different session augmentation approaches and weighting ways. In final, the system is boosted at most with the session augmentation consisting of unique IDs decomposition, context window-based history generation, and extracting identical hierarchy. The average of precision, recall and F1 measure has been used as performance metric with different cross validation settings and it's found that female prediction accuracy is higher than the one for male prediction . In another research, a method is proposed to predict the gender information of each customer by using their transaction data on e-commerce systems, such as time of access, products and list of categories displayed [1]. A machine learning approach is done in this study and Random Forest and Support Vector Machines(SVMs) classifiers are compared with each other by considering different supporting techniques for handling class-imbalance issue such as cost-sensitive learning, resampling and class balancing. As a result, the best results are obtained with Random Forest Classifier built by using cost-sensitive learning with class balancing. The results are compared in terms of balanced accuracy and macro F1 score. The best model had 81.2% on balanced accuracy and 81.4% on macro F1 score in this research. An additional study exists for the prediction of the genders of customers [6]. This study uses product viewing logs of users and it presents two components in the architecture created for gender prediction, which are machine learning model and label updating function. Machine learning algorithm learns a classification model by using the behavior, temporal and viewed product features, after genders of new users in the test set are predicted, the second component label updating function is being used. This component uses product information to revise the labels of test data that are learnt by a machine learning algorithm. In this study, the best model is obtained with Gradient Boosting Decision Trees(GBDT) and to solve imbalanced data issue, it is used with cost-sensitive learning and sampling techniques such as down-sampling or over-sampling. The results have shown that the introduced method works better for female prediction.

2 Approach

In this project, train and test transaction data sets are provided and it is aimed to have a model that predicts gender information of the customers accurately. At first, a data set named all data, including both train and test sets, is created to practice feature engineering in a single framework. After feature engineering steps are completed, learning of a classification model is held and test results are compared by using the provided online excel spread sheet in terms of their AUC and BAR values. The best performance is obtained with the model obtained by using Light Gradient Boosting Machine Classifier, which has an AUC value of 0.883 and a BAR value of 0.826, meaning

0.854 as a final performance result on the test data. Approach to this project will be explained in detail in the following parts: Data Preparation, Modelling and Evaluation.

2.1 Data Preparation

After all data, which is mentioned above, has been obtained, manipulations started to be made. At first, action period feature holding the day period information of each transaction with four different values morning, noon, evening and night, was extracted from the time stamp feature. Additionally, weekend weekday feature that gives information about whether the transaction is occurred in weekend or weekday was obtained by the time stamp information.

In the next step, since the all data includes missing product gender information which has a potential to be critical feature for the user gender prediction, a random forest classifier that predicts product gender has been developed. In order to have high accuracy in this model, some features are created based on content id. These features are normalized values of each user action, average selling price of the product corresponds to that content id, normalized values of genders observed in that content id, normalized values of weekend weekday observations of that content id, normalized values of action period observations corresponding to that content id, number of records of that particular content id and one-hot encoded level1 category names. Finally, target of this data containing 610.035 instances and 25 features is product gender. The instances with filled product gender info are used as training set whereas rest of it, which is trying to be filled, is used as test set. Random Forest Classifier is learnt on training set with an average balanced accuracy of 0.72 obtained by 5-fold cross validation. By using this model, missing product gender information of the data set, which is named as all data, are filled.

After the product gender information is filled in all data having 2.955.345 instances, the feature engineering steps continued with the creation of new features. Firstly, for each action type, product genders with normalized values, action periods and weekend weekday information with normalized values, average, minimum and maximum selling prices, number of observations, level 1 categories with their normalized values are obtained as separate features after grouping with respect to unique ids. Also, by using the average selling price observed in the transactions of particular user in particular action, segment features are obtained for each action with low, mid and high values. Later, number of observations in total and number of observations in each action for each unique id are obtained as features. After the implementations are made in all data for each unique id, the data set appeared to have 7998 instances with 127 features. This data including the features of each unique id is divided into known and submission features with respect to unique ids of train and test sets. Furthermore, known features are split into two as train and validation sets for the evaluation purposes.

Before finalizing the train, validation and test sets, a final implementation is made by using all data created at the start. This final implementation aims to extract the information lost at aggregation from long format data After separating all data into train, test and validation data sets by using the corresponding unique ids determined just before this step, a Random Forest Classifier is fitted in train data, which includes 1.153.786 instances and 204 features, to predict gender for each transaction. By using out of bag (OOB) decision function to obtain OOB predictions of training data, and using the fitted model to obtain predictions of validation and test features, 4 new features are obtained. These features are minimum, maximum, mean and standard deviation of predicted probability of being female for each unique id. After the OOB features obtained from long data

with a random forest model, they are merged to corresponding train, validation and test data sets. At last, modeling step is started with train set containing 4494 instances and validation set containing 1124 instances with 131 features. On the other hand, some trials are also made by training data that does not contain all data training or OOB features. Yet, as will be mentioned in the results part, the best performance is obtained with the data set explained just above in detail.

2.2 Modelling, Tuning& Evaluation

In the modeling phase, LightGBM classifier and Random Forest classifier have been trained with several data sets, as will be explained with its details in Results part. Since both random forest and LightGBM has large number of hyperparameters a tuning scheme is needed to obtain better performances. For Random Forest random grid search is employed with a few number of iterations. The grid of the random forest included number of estimators, min samples split, min samples leaf and max features. Also class weight is set to balanced since the number of females are much more. For LightGBM a hyperparameter tuning library called optuna is employed. Optuna uses a smart sampling heuristic to make search in given grids. In order to tune the boosting model a search for at least 20 minutes and at most 1 hour is made. The search space included number of estimators, max depth, bagging fraction, feature fraction, min child samples and more. During the search mean of balanced accuracy and AUC score is checked and the parameters that obtained the best result in 5 fold cross validation is returned.

The best performance is observed with the final data set explained above in detail with LightGBM classifier. LightGBM has outweighed Random Forest in terms of robustness, test performance and submission performance that is evaluated in the leader board system. AUC and balanced accuracy score are used as performance metrics and their average is taken as the overall test performance.

3 Results & Discussion

The results in this task is obtained from two different evaluation sets. First the test data that is separated from the non submission data. Second a part of submission data that is evaluated by the leader board system. In order to decide which model to choose both performances are evaluated and one is chose by considering the values of both evaluations to have robust final submission. The table below shows important trials made during the period and how they performed. It can also be a summary of how the approach is evolved through time since it is sorted and it also includes the properties of the approach.

Final Prediction Algorithm	Product Gender Fill Method	Categorical Avg Price	Long Data Training	OOB Features	Tuning Method	Test Performance	Summission Performance
Random Forest	Category Based	FALSE	FALSE	FALSE	Random Grid Search	**0.8278**	**0.8375**
LightGBM	Category Based	FALSE	FALSE	FALSE	Optuna 5min	**0.81993**	**0.8254**
Random Forest	Category Based	TRUE	FALSE	FALSE	Random Grid Search	**0.8231**	**0.8407**
LightGBM	Category Based	TRUE	FALSE	FALSE	Optuna 20min	**0.8336**	**0.8472**
LightGBM	Category Based	TRUE	OOB Features	Mean	Old Optuna 20min	**0.8329**	**0.8484**
LightGBM	Category Based	TRUE	OOB Features	Mean Min Max Std	Optuna 30min	**0.8313**	**0.8547**
LightGBM	Random Forest Model	TRUE	OOB Features	Mean Min Max Std	Optuna 1h	**0.8362**	**0.8547**
Ensemble of OOB and no-OOB LGBM models	Random Forest Model	TRUE	OOB Features	Mean Min Max Std	Optuna 20min	**0.8371**	**0.8491**

Figure 2: Results obtained from different trials.

During the trials, it is noticed that LightGBM is capable of reaching better performance values than random forest with a little tuning with the same feature set as row 3 and 4 shows. However, random forest is still employed in OOB feature extraction for its OOB prediction property and for product gender fill since it has lower hyperparameter complexity. Another observation was using avg price information as categorical helps to get better performances most probably due to not allowing to overfitting to that feature. The other jump in the data is made with introduction of OOB features from long data training. There is information in the long data that does not represented in the aggregated data. Therefore OOB features helped models to employ different information for better performance. Since OOB features are very effective on model and changes the whole training procedure, the effect of them has seen after a proper tuning. It is also possible to say that filling product gender with a better performing method helped a little. Finally, ensemble of models from data with OOB features and data without OOB features did not work well. It is believed that this result is due to caching similar information from data.

As a result the model marked with green is chosen as the final submission due to its both high test performance and submission performance.

4 Conclusion & Future Work

In this project a gender classification problem is studied. First an unstructured dataset is obtained and missing values are filled. Some features are processed and new features are added to the dataset. Since the problem is a variation of multiple instance learning problem, bag representations of the whole data is created. Different methods are tried to learn from the training dataset. Best performing method ,gives AUC value of 0.883 and a BAR value of 0.826, is selected as the submission method.

Along the modeling part it is observed that product gender feature has the most importance among all features. Since it has missing values in the raw data, better estimation of missing product gender values will yield better result. The estimation of missing product gender values could be studied longer and a better estimation method could be implemented to improve the result.

In this project, only random forest and LightGBM methods are studied. Other learning methods such as, probabilistic versions of SVM and KNN can be used. Since it is a multiple instance

learning problem, the related problems in the literature could be searched and be implemented in this project.

Some unique users have fewer appearance in the data set. The results comes from those users could be misleading and they can be studied separately from the rest of the data. Learning methods that performs better with smaller datasets could be used to improve the overall result.

5 Code

https://github.com/harunkivril/IE582-TermProject

References

[1] Duong Tran Duc, Pham Bao Son, Tan Hanh, et al. A resamping approach for customer gender prediction based on e-commerce data. *Journal of Science and Technology: Issue on Information and Communications Technology*, 3(1):76–81, 2017.

[2] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[3] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[5] Mohammad Masud Khan, Mohammad Golam Sohrab, and Mohammad Abu Yousuf. Customer gender prediction system on hierarchical e-commerce data. *Beni-Suef University Journal of Basic and Applied Sciences*, 9(1):1–12, 2020.

[6] Siyu Lu, Meng Zhao, Hui Zhang, Chen Zhang, Wei Wang, and Hao Wang. Genderpredictor: a method to predict gender of customers from e-commerce website. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 13–16. IEEE, 2015.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Yoo-Kyoung Seock and Lauren R Bailey. The influence of college students' shopping orientations and gender differences on online information searches and purchase behaviours. *International Journal of Consumer Studies*, 32(2):113–121, 2008.