Fancy Title

Business Analytics and Data Science Group Project

submitted to

First Prof
Second Prof

Humboldt-Universität zu Berlin
School of Business and Economics

Chair

by

Claudia Guenther, Phi Nguyen, Julian Winkel
Immatriculation Numbers

Anything else we want to say

Berlin, Date

Abstract

Insert abstract here

List of Tables

List of Figures

Abbreviations

ANN	Artificial Neural Network
LM	Linear Model

Contents

1 Introduction

2 Previous Literature

1 Page

Divide previous research in subsections that will be presented in the following.

This is how we cite Badea (2014). The reference is automatically pasted in the according section. You can also cite indirectly at the end of a sentence (Badea 2014). In this format, it is possible to insert pages, too (Badea 2014, 10–14).

3 Methodoloy

2 pages

3.1 Heterogeneous ensemble models

The idea behind ensemble models (EM) is to combine multiple learning algorithms that taken together improve prediction accuracy over the performance of the individual models. The success of EM in predictive analytics is due to their ability of reducing both bias and variance by combining multiple model forecasts (Dietterich 2002). EM can be divided into homogeneous and heterogeneous approaches. Homogeneous EM use a single base-learning algorithm (e.g. decision tree) on distinct training sets. In contrast, heterogeneous EM train and combine learning algorithms of different types. In the context of classification problems, heterogeneous EM are also referred to as multi-classifier systems (MCS) and have gained increasing popularity (Ranawana and Palade 2006). Two key design features for the success of MCS is the diversity among the selected classifiers as well as the good individual performances (strength) of the baseline classifiers (Ali, Sremath Tirumala, and Sarrafzadeh 2015). However, is has been shown that neither diversity nor individual performance of baseline models can directly predict the ensemble success (Ranawana and Palade 2006).

MCS have been shown to be useful in predicting customer e-commerce behavior, outperforming single classifiers in prediction accuracy (Tsoumakas, Partalas, and Vlahavas (2008), E. Kim, Kim, and Lee (2003)). Specifically, Heilig et al. (2016) demonstrate the suitability of an ensemble selection approach (ES) for predicting product returns in online retailing. This three-step ensemble technique builds upon the creation of a library of heterogeneous candidate base models. Each classifier method can be used to build up multiple models with different meta-parameter settings. The ensemble model is created by combining base model from the library through an extensive search strategy. Though this approach has been demonstrated to be very effective for accurate predictions, it nevertheless requires extensive computational time. We therefore adopt a standard two-step ensemble approach. First, we develop and train four base classifiers models. Second, we combine these individual models with a simple non-linear combination method, namely majority voting. In case of a voting tie, the base model with the highest single accuracy will act as a tiebreaker.

3.2 Base classifier models

- baseline models can be seen in table 1
- in how far are they diverse? why did we select those?
- which one can capture linear, non-linear effects?

4 Data

4.1 Data sets

The two data sets available to us contain a total of 150,000 order records from an online apparel retailer from a yearlong selling period. For 50,000 of these records it is unknown whether an ordered item has been sent back by the customer or not. This second data set is the subject of our binary predictions of customer's returning behavior (return/not return). Both data sets include a total of 13 continuous and categorical variables. These covariates give information on customer demographics (e.g. user state, date of birth, title), order details (e.g. order date, delivery date), and item characteristics (e.g. item size, price or color). To prepare the data sets for our analysis we apply a set of standard pre-processing actions. Following the careful inspection of each variable, we remove all implausible values (e.g. extreme outliers). Approximately 20% of all records have missing values in either the *delivery date* and *date of birth*. For bettter comprehensibility, we transform these variables *delivery time* and *age* respectively. Before imputing missing values, we flag them using dummy variables to extract their predictive power (Lessmann 2016, 18). Since *age* seems to be missing (completely) at random (MCAR) according to our data inspection, imputing it using mean substitution gives us an unbiased estimates (Schafer and Graham 2002)[1].

Missing values in delivery time, caused by missing delivery dates, are clearly not missing not at random (MNAR) as they have a zero mean return rate and therefore are a perfect predictor. Possible reasons for these missing values are manifold. Without knowing the process generating these MNAR values, we cannot find unbiased substitutes form them (Schafer and Graham 2002, 171). We adopt three single substitution methods, namely case dropping, mean and median imputation, and chose the latter one based on model performance. We standardize the continuous variables only after the feature creation step to maintain their interpretability. Likewise, we do not directly drop zero (almost) zero variance predictors (e.g. date of birth) since we use them for feature extraction.

5 Feature engineering

5.1 Feature construction

Comprehensive feature engineering, i.e. creating and selecting useful and informative features, is the key to successful e-commerce prediction models (G. Liu et al. 2016). Based on the data structure available we separate our indicators into three e-commerce-related categories, namely product, customer and basket related features. Based on the data structure available we separate

[1] Additionally, we carry out a Maximum Likelihood imputation of age in case the missing values are only missing at random (MAR), yielding the same model performance.

our indicators into three e-commerce-related categories, namely product, customer and basket related features. Features in the product category are all item-specific indicators (e.g. item size, item price). Likewise, customer-related features contain variables with relational or demographic information on each customer (e.g. customer age, age of user account). We define the basket-related features as those that are common within one order basket (e.g. number of items within a basket). We consider all items belonging to one basket that were ordered by the same user on the same day. In order to extract as much information as possible from the 13 original variables, we apply a variety of different extraction and projections methods. This includes unsupervised methods (e.g. equal frequency binning) as well as supervised approaches, like mixed clustering (hierarchical and k-means). Our extraction methods can be broadly summarized in four groups: discretization, projection, transformation/combination of variables and data augmentation by using external information. Based on these methods we construct an enriched data set with a total of 45 features. We apply discretization to several continuous variables, as this can be useful for capturing non-linear effects, constructing a more comprehensible representation of variables or improving predictive accuracy, especially in the context of tree-based decision algorithms (H. Liu et al. 2002). For example, we construct price item bins using equal frequency discretization as we suspect non-linearity regarding the target variable. This approach is superior to equal width discretization for our purpose since it is outlier-resistant and gives us a higher interval resolution for item prices with a high density (Cichosz 2014, 531).

A special challenge of this data set, which applies to many predictive modeling settings on customer behavior, is the presence of high-cardinality attributes (e.g. user id, item id). We replace each category with its Weight of evidence (WOE), a numerical projection method for categorical variables. The WOE is a measure of predictive power of a category with respect to the target variable, which has gained popularity in credit scoring (Lessmann 2016, 19–20). For a given customer, it shows a customers tendency in returning behavior, a valuable predictor variable. Replacing high-cardinality attributes with their respective WOE also reduces the dimensionality of our data set drastically. Since WOE has been shown to work well across the different types of our baseline models and requires low computational effort, we apply this projection methods to all categorical variables(Lessmann 2016, 32–33). For categories where WOE is not computable due to the existence of empty classes, we adapt WOE calculation by adding artificial data points as suggested by Kohavi and John (1997).

Another useful approach we take is the transformation and combination of variables based on domain knowledge about consumer's shopping and returns behavior (Arnold and Reynolds (2003)). For example, we expect that discounted items will have a lower return rate due to lower item quality expectation as proposed by Petersen and Kumar (2009). Thus, we construct several discount variables (absolute discount, percentage discount, discount dummy) by taking the maximum item price for each item size as proxy for the regular item price and considering all deviations as a discount. Furthermore, a common defect of online shopping is that customers might be unsure about which size or color they should order based on the item description and thus place multiple orders of the same item (Foscht et al. 2013). We therefore construct informative variables on the order basket of consumers, e.g. the basket size in terms of items or the number of similar items with distinct sizes within one basket. Furthermore, we augment and enrich the data sets by including external sources to construct new features. For example, we use statistics on the mean income for different age groups and different regions within Germany. By combining these indicators with each customer's demographical data (age and user state) we are able construct a proxy for each customer's income. Moreover, we tackle the lack of information on item categories (e.g. pants, shoes, clothing, accessories). Based on size tables for different apparel categories we gathered we are able to categorize over 90% of items by finding sizes unique to a category. As expected, the mean return rate of shoes is almost twice as high as mean returns in accessories.

5.2 Feature selection

Finding an optimal feature subset is important for various reasons: First, learning algorithms like neural networks or tree-based models have been shown to degrade in performance when confronted with datasets containing redundant or irrelevant variables (Zdravevski et al. 2014). Second, feature selection reduces data dimensionality, reducing the needed computational power for model building process and improving model comprehensibility. We adopt a hybrid feature selection strategy by combining filters and wrappers. As a first preprocessing step as proposed by Zdravevski et al. (2014), we remove very poor variables based on model agnostic information criteria. This has the advantage of reducing space dimensionality quickly and address overfitting (Guyon and Elisseeff 2003, 1170). For numerical variables we compute the F-Score as proposed by Y.-W. Chen and Lin (2006), whereas the average information value (obtained from WOE) is computed for categorical covariates. The second step for our feature selection is sequential backward elimination wrapper. This Greedy selection strategy is advantageous since it has been shown to be robust against overfitting and is designed to optimize model performance (Guyon and Elisseeff 2003, 1167). Since the baseline models from our ensemble might perform best on distinct feature subsets, the optimal approach would be to conduct an individual wrapper for each model. Due to computational constraints we are nevertheless limited to conducting a single wrapper for all models. Since feature subsetting is especially important for tree-based models and neural networks, we design a wrapper based on a simple random forest. First, we create a hold-out test set by creating a random stratified test and train data set. After standardization of both data sets, the WOE of categorical variables is calculated on the training set and then passed on to the test set to prevent overfitting. Finally, we remove features by means of sequential backward elimination based on maximizing AUC. Our final subset contains only XX features, which gives us an mean AUC of 0.728 on the test set.

6 Model building

6.1 Experimental design

6.1.1 Baseline models

6.1.2 Candidate selection and combination

6.2 Performance Measurement

- discuss AUC, accuracy, costs
- post-processing

7 Results

8 Conclusion

9 References

Ali, Shahid, Sreenivas Sremath Tirumala, and Abdolhossein Sarrafzadeh. 2015. "Ensemble Learning Methods for Decision Making: Status and Future Prospects."

Arnold, Mark J, and Kristy E Reynolds. 2003. "Hedonic Shopping Motivations." *Journal of Retailing* 79 (2). Elsevier: 77–95.

Badea, Laura Maria. 2014. "Predicting Consumer Behavior with Artificial Neural Networks." *Procedia Economics and Finance* 15: 238–46. doi:https://doi.org/10.1016/S2212-5671(14)00492-4.

Chen, Yi-Wei, and Chih-Jen Lin. 2006. "Combining Svms with Various Feature Selection Strategies." In *Feature Extraction*, 315–24. Springer.

Cichosz, Pawel. 2014. *Data Mining Algorithms: Explained Using R*. John Wiley & Sons.

Dietterich, Thomas G. 2002. "Ensemble Learning." *The Handbook of Brain Theory and Neural Networks* 2. MIT Press: Cambridge, MA: 110–25.

Foscht, Thomas, Karin Ernstreiter, Cesar Maloles III, Indrajit Sinha, and Bernhard Swoboda. 2013. "Retaining or Returning? Some Insights for a Better Understanding of Return Behaviour." *International Journal of Retail & Distribution Management* 41 (2). Emerald Group Publishing Limited: 113–34.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3 (Mar): 1157–82.

Heilig, Leonard, Julien Hofer, Stefan Lessmann, and Stefan Voß. 2016. "Data-Driven Product Returns Prediction: A Cloud-Based Ensemble Selection Approach." In *ECIS*, Research–in.

Kim, Eunju, Wooju Kim, and Yillbyung Lee. 2003. "Combination of Multiple Classifiers for the Customer's Purchase Behavior Prediction." *Decision Support Systems* 34 (2). Elsevier: 167–75.

Kohavi, Ron, and George H John. 1997. "Wrappers for Feature Subset Selection." *Artificial Intelligence* 97 (1-2). Elsevier: 273–324.

Lessmann, Stefan. 2016. "Data Preprocessing in Database Marketing: Tasks, Techniques, and Why." In *Advanced Database Marketing: Innovative Methodologies and Applications for Managing Customer Relationships.* Routledge.

Liu, Guimei, Tam T Nguyen, Gang Zhao, Wei Zha, Jianbo Yang, Jianneng Cao, Min Wu, Peilin Zhao, and Wei Chen. 2016. "Repeat Buyer Prediction for E-Commerce." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 155–64. ACM.

Liu, Huan, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. 2002. "Discretization: An

Enabling Technique.” *Data Mining and Knowledge Discovery* 6 (4). Springer: 393–423.

Petersen, J Andrew, and V Kumar. 2009. “Are Product Returns a Necessary Evil? Antecedents and Consequences.” *Journal of Marketing* 73 (3). American Marketing Association: 35–51.

Ranawana, Romesh, and Vasile Palade. 2006. “Multi-Classifier Systems: Review and a Roadmap for Developers.” *International Journal of Hybrid Intelligent Systems* 3 (1). IOS Press: 35–61.

Schafer, Joseph L, and John W Graham. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7 (2). American Psychological Association: 147.

Tsoumakas, Grigorios, Ioannis Partalas, and Ioannis Vlahavas. 2008. “A Taxonomy and Short Review of Ensemble Selection.” In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications.*

Zdravevski, Eftim, Petre Lameski, Andrea Kulakov, and Dejan Gjorgjevikj. 2014. “Feature Selection and Allocation to Diverse Subsets for Multi-Label Learning Problems with Large Datasets.” In *Computer Science and Information Systems (Fedcsis), 2014 Federated Conference on*, 387–94. IEEE.

Declaration of Authorship

TEXT

15.01.2018