Fancy Title

Business Analytics and Data Science Group Project

submitted to

First Prof
Second Prof

Humboldt-Universit?t 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 Predictive Analytics or other title

3.2 Ensemble Models

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). We standardize the continuous variables only after the feature creation to maintain their interpretability. 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. 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 this 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.

4.2 Feature creation and selection

[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.

5 Model building

5.1 Experimental design

5.1.1 Baseline models

5.1.2 Candidate selection and combination

5.2 Performance Measurement

- discuss AUC, accuracy, costs
- post-processing

6 Results

7 Conclusion

8 References

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.

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.

Declaration of Authorship

TEXT

15.01.2018