Business Analytics and Data Science
Term paper

Thomas Siskos, 580726

February 24, 2019

1 Introduction

During online purchases customers often send back items they order. These returns are costly and because of the high competition in online retail it is not possible or highly inadvisible to pass on the costs of return shipping to the customer. Therefore, accurate predictions of product returns could allow online retailers to impede problematic transactions, for example by restricting payment options or by displaying a warning message and thus cut down on cost due to shipping.

Section 2 contains a description of the data and the exploratory results results. Section 3 describes the actions taken in order to clean the data and a brief overview of the efforts taken during feature engineering. Section 4 specifies the different algorithms that were deployed and section 5 includes a succinct discussion of the results. Section 6 tries to minimize the costs of misclassifying an item directly by using a custom built and modified Genetic Algorithm. Finally, section 7 contains concluding remarks.

2 Exploratory Data Analysis

3 Data Preparation

The data contained numerous missing values. Some of these were obscured, specifically for the delivery date it seemed that dates lying as far back as 1994 were being used to encode a missing value. However, once encountered, these missing values were easy to impute by the mean number of days passed between the order and the delivery of the item for cases were the delivery date was available. When looking at the distribution of days between order and delivery it seemed more in order to use the median since it was heavily skewed, yet imputing by the mean seemed provide better results. Similarly, for some users it was difficult to compute their age, since they either did not provide their day of birth or instead opted to provide implausible ones. Consequently, all years of birth lying farther back than 1926 were removed and the age of these users

Table 1: Selection of engineered features

	feature	description
users	`tenure`	days between registration and order
items	`price-off`	discount compared to maximum item price
orders	`num items`	count of item IDs in order
	`days until delivery`	days between order and delivery
	`num sizes`	count of unique sizes
	`total value`	sum of all item prices in order
	`num colors`	count of unique colors
	`seq number`	enumerate order date per user
brands	`brand mean price`	average price of item's brand
state	`state mean delivery`	average number of days until delivery

was imputed by the difference in days between registration date and the birth date, of the valid users which was subtracted from the registration date of the incredulous ones.

The data contained a large number of categorical variables which in turn contained numerous levels. Especially the items' colors involved some spelling mistakes and extravagant names for different shades of the same color. Both problems were solved by manually sifting through the various labels and summarizing the more detailed color names into broader categories. This way it was able to reduce the 85 initial colors to 14 unique levels in the cleaned data. For these densely populated categories it was possible to calculate the Weight of Evidence (Gough, 2007).

The items' sizes proved to be difficult to clean. Ideally one would want to extract categories like *small*, *medium*, *large* and while these are provided in some, they are not provided in the majority of cases. Instead there is a whole clutter of different sizes and without knowing the type of clothing only limited information can be extracted. Through the use of regular expressions, for example, it is possible to determine if items are pants, since they have exactly four numerical digits (two for the width and two for the length).

Table (1) contains a selection of the most important engineered features. Furthermore, all possible pairwise ratios and interaction terms were computed where the most correlated features were discarded afterwards.

Figure 1: Feature Correlation Plot

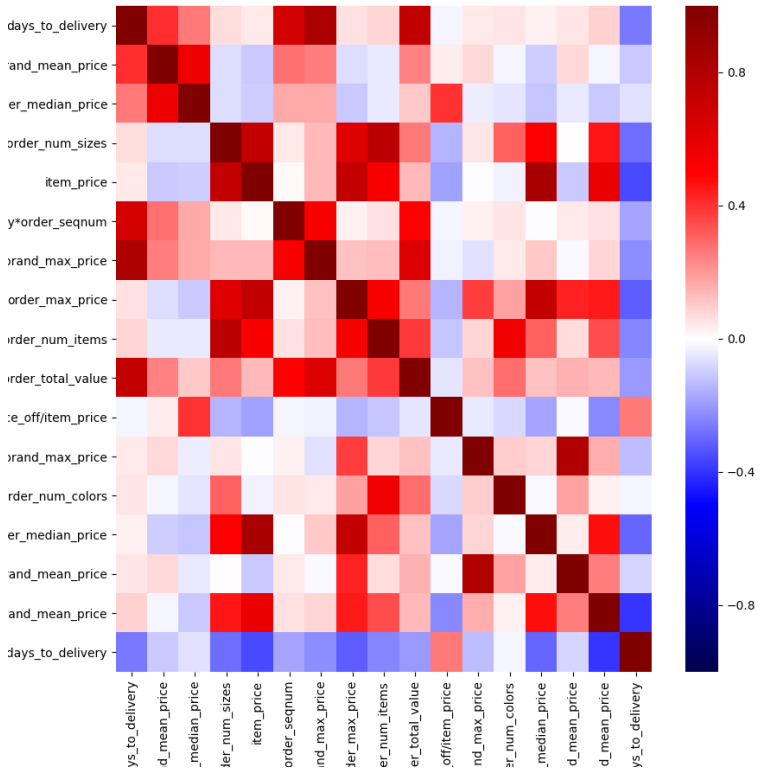

Table 2: Selected Results

	Train AUC	Test AUC
Feed forward Neural Network	0.711	0.705
Random Forest	0.718	0.704
Gradient Boosted Trees	0.773	0.712

4 Model Tuning and Selection

5 Model Evaluation

6 Minimizing Costs directly

7 Conclusion

References

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213-228. Retrieved from `https://doi.org/10.1080/02671520701296189` doi: 10.1080/02671520701296189