Data Science report on Black Friday sales

Florian Störtz

We thank you for your trust and confidence in this matter, and in the following would like to present the insights we have gained in a first investigation into your Black Friday sales dataset.

1 Why Data Science?

Our investigation goes far beyond what is achievable by a mere descriptive statistics approach. The dataset with which your company has supplied us contains a great set of meaningful features, but it is in the interaction between them that contains the insight into purchasing behaviour. For this, we employ sophisticated data science tools (see below) which require specialist design and interpretation. We are glad to be able to provide both.

2 Descriptive Analysis of the Dataset

In this section, we want to conduct a surface analysis of the data. As can be seen from figures 1 and 2, most purchases are made by 26-35 year olds, where unmarried users contribute the larger part of the buyer group, especially among female customers. Occupations 0, 4 and 7 are responsible for the most purchases by occupation.

The number of purchases done by each individual user shows similar results (figure 3).

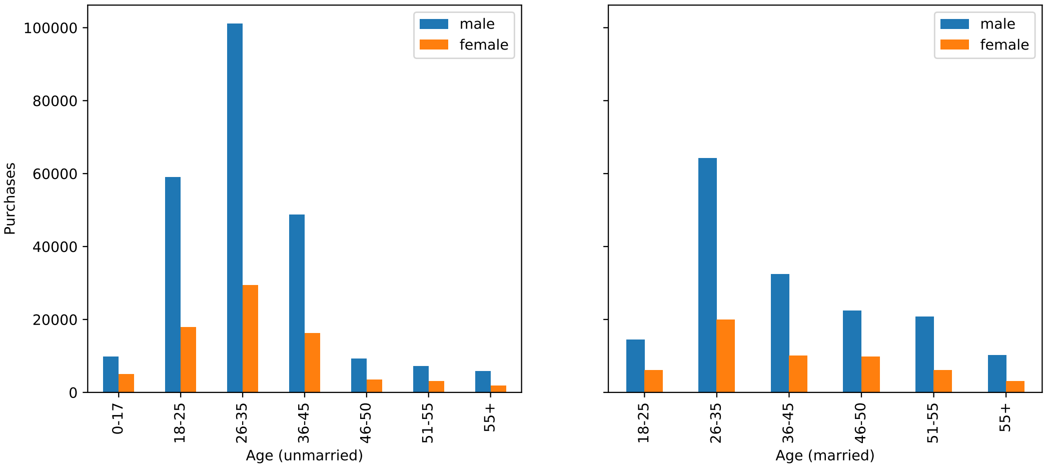

Figure 1: Number of purchases by age group.

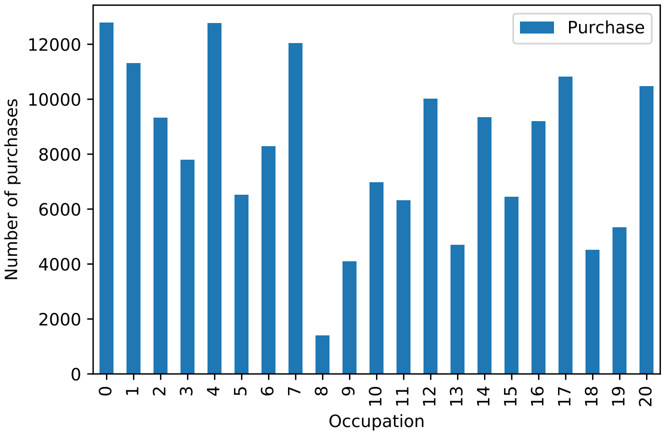

Figure 2: Number of purchases by occupation.

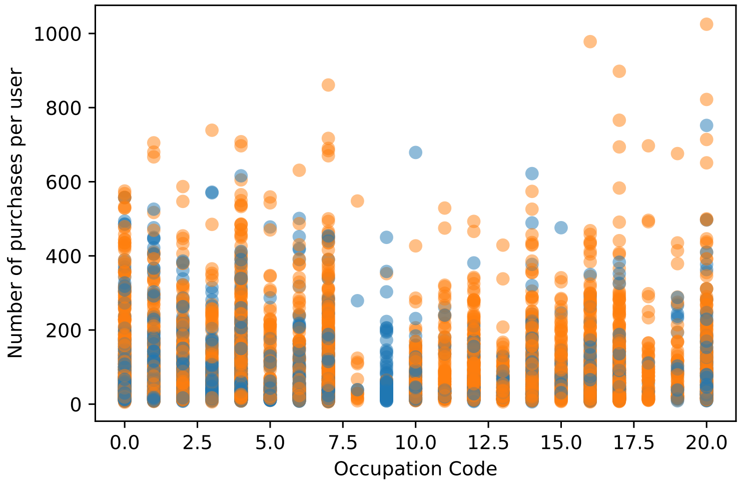

Figure 3: Number of purchases per user, given their occupation. Blue dots signify male users, orange dots signify female users. It is interesting to observe that occupation 9 is dominated by male users.

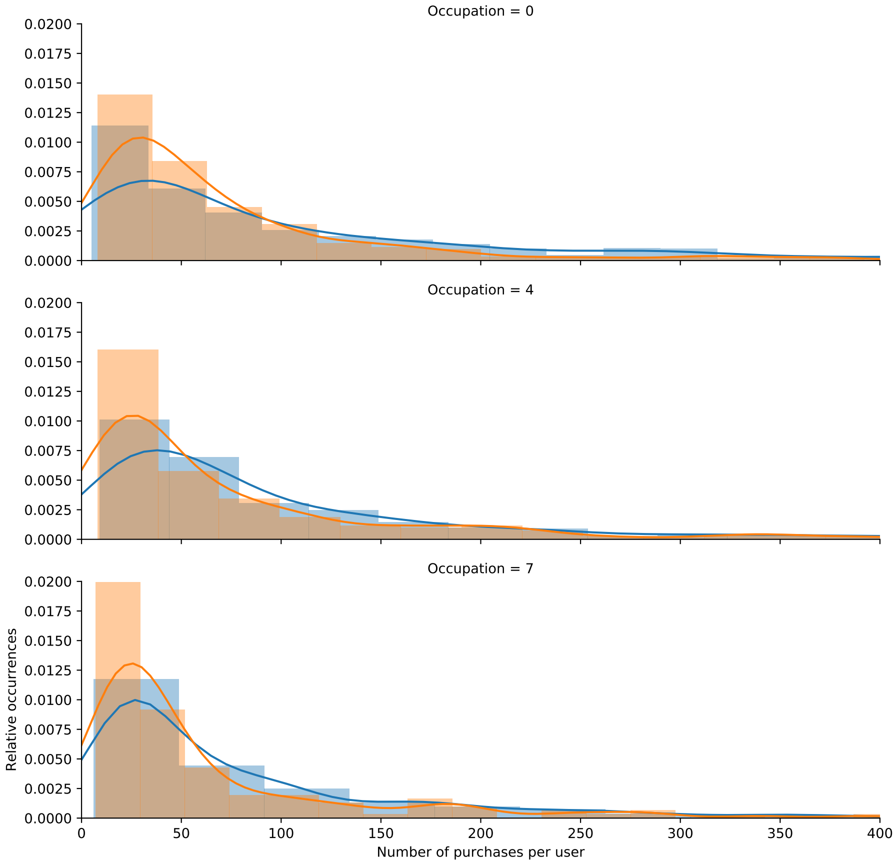

Figure 4: Distributions of the number of purchases per user for the three key occupation groups found in figure 2. We can see that the distributions extend farther to the right for male customers (blue) than for female customers (orange).

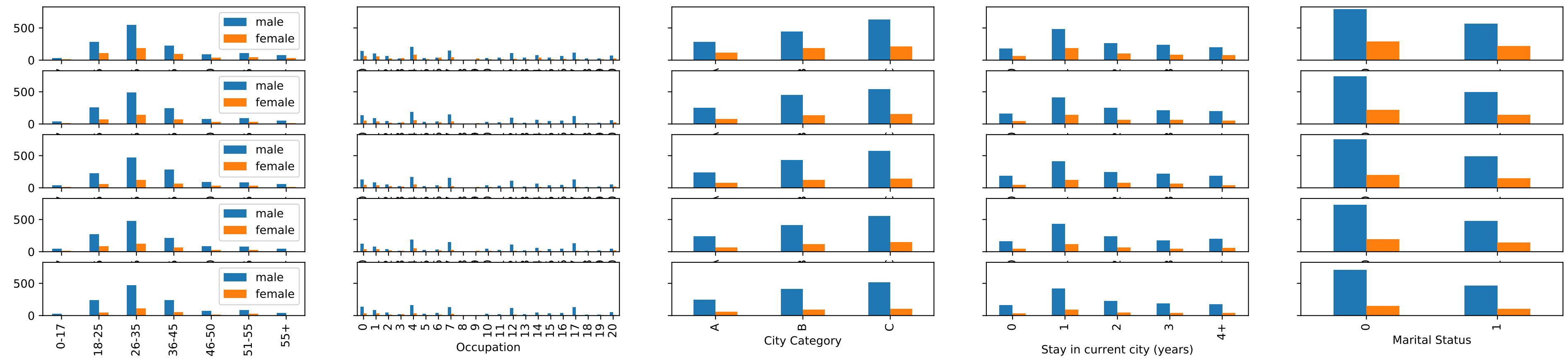

Figure 5: Number of purchases for the five best-selling products, organised by feature. In decreasing order of purchases, from top to bottom row, these products are: P00265242, P00110742, P00025442, P00112142, P00057642.

Figure 5 shows that products P00265242, P00110742, P00025442, P00112142, P00057642 are the best-selling. One can begin to gain insights into the distribution of purchases along the supplied features, such as the prevalence of occupations 0, 4, 7 and 12 or of the 26-35 age group, but for the reasons below, we should leave these conclusions to the next section. Feature refers to the columns of the supplied data file, except for the User_ID, Product_ID and Purchase columns.

3 Machine Learning Analysis of the Dataset

We use a highly efficient and flexible gradient boost framework, XGBoost [1], in order to extract knowledge about which factors influence purchasing behaviour which is hidden in the dataset. This approach uses a combination of established sub-models, each correcting for the errors made by the others. More sub-models are added until no further improvement to the prediction performance can be made.

In short, we feed the model a combination of real purchases from the dataset and fake ones (i.e. a random combination of features), and give the model the objective to distinguish between the two. The internal parameters of the model are continuously adjusted so as to get closer and closer to this aim.

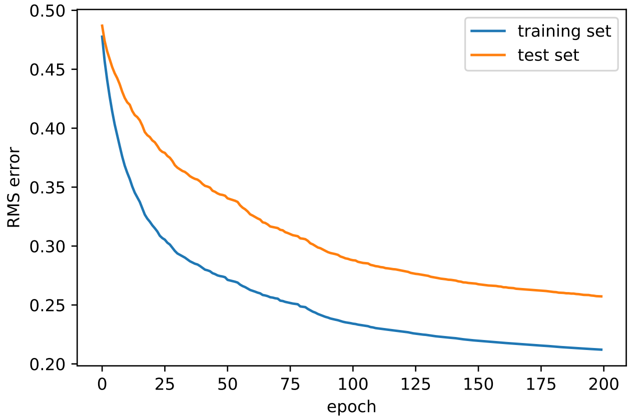

Figure 6: Training performance of our gradient boost classifier, on the actual training set (blue) and on unseen test data (orange) which make up for 80% and 20% of the whole data set, respectively. This behaviour is to be expected of a classifier of decent performance.

Given the learning data in figure 6, we can see that the model behaves as expected – the loss on the training set goes down roughly exponentially with the number of optimisations (blue curve); the tendency is mirrored by the loss on an unseen portion of the data (orange curve). This proves that the network indeed learns to differentiate between real and randomly generated purchases, and can thereby pick up on what real purchases depend on. The fact that the errors diverge indicates overfitting, probably due to the fact that the tree structure which xgboost uses is not adjusted to the structure of the data. Learning becomes slower and slower, such that we do not achieve a perfect predictor.

In order to then extract a human-readable indicator for the importance of the single features, i.e. whether and how much each single feature contributes to the decision of purchase or not, we use an existing framework to calculate Shapley values [2]. Shapley values of the twenty most influential features are depicted in figure 8.

The advantage of this approach is that it takes into account interactions between the single features, which cannot be scrutinised from figure 5, e.g. when features are confounded by the interaction of others.

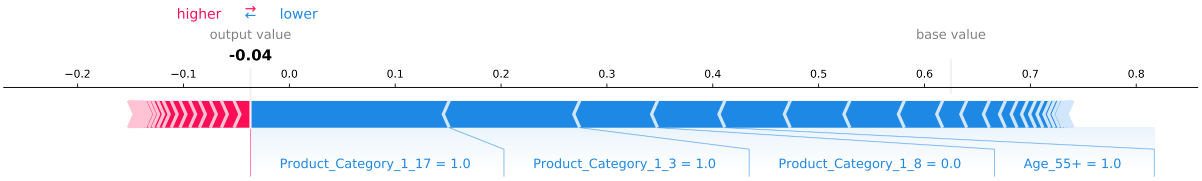

Figure 7: Example of how the single feature values influence the prediction whether a purchase is deemed to be real (final value above 0.5) or not (final value below 0.5). This particular data point will be classified as not a real purchase.

Figure 8 shows that the prediction of our model is first and foremost dependent on product categories. Categories 5 and 8 seem to be conducive to a purchase, while categories 18, 7 and 12 are not. The next important feature is gender; the model again supports the notion that male customers (blue dots to the right of the y-axis) are more likely to purchase. Age groups are also an important feature; the red dots to the left of the y-axis for age group 0-17 indicates that this is not our target group for Black Friday sales. Contrary to figure 5, marriage status is only a weak indicator for purchase. This means that this

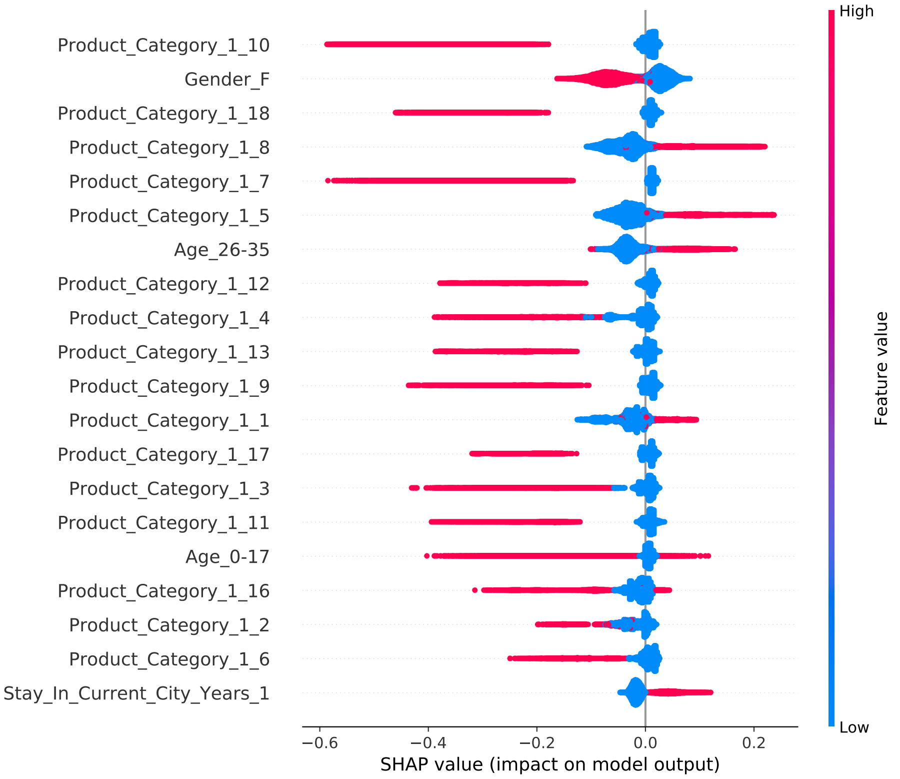

Figure 8: Shapley feature importance values of the 20 most influential features. Values to the right of the y-axis indicate that the specific feature is connected with increased sales; to the left they indicate the opposite. Red dots indicate that the feature is set for the given data point, blue value indicate that it is not set.

factor seems to be confounded by other features. It is worth noting that the group of customers who has stayed in the city for around one year is most prone to purchase.

The above analysis is reliable insofar as it assumes that real purchases only depend on the features given in the dataset. The model performance depicted in figure 6 is no reason for concern, but nevertheless shows that a classification algorithm which is more suited to the structure of the data, e.g. a bespoke neural network architecture, could yield better predictions and lower the error even further.

4 Strategic Advice

Taking all these results into account, a sales strategy for the upcoming Black Friday still depends on the perceived elasticity of demand in customers. Assuming that the market is not fully saturated yet and demand is still elastic, a sales strategy could look as follows:

- Female and married customers should be increasingly addressed.
- Products of categories 1, 5 and 8 should be stocked the most; products of categories 7, 10 and 18 should be reconsidered.
- The target audience, which at the moment centres around 26-35 year olds, could be extended.

In order to come up with more meaningful predictions, it would be desirable to collect features which are not confounded, i.e. depend on the interaction of underlying, hidden features such as we have seen above. These could be features about the products itself (more than a coarse category) and about buyers. The investigation above could be extended by collating individual buyer's behaviour, which given the limited number of purchases per user in the given dataset was not possible.

References

[1] Chen *et al.* arXiv:1603.02754v3 (2016).

[2] Štrumbelj *et al.* Knowl. Inf. Syst. **41**, 647 (2014).