Predicting the number of new paying customers in the next month

Matej Miočić[1]

Abstract

Having the information of how many new paying customers will a company gain in the next month has a direct impact on their revenue. Databox is storing customers' information about their event data. With this data combined with number of paying customers from previous months we try to predict the number of next month's paying customers. With the data we have, we show that our models outperform the majority classifier by approximately 5%.

Keywords

Databox, paying customer, predictive modeling, events

[1] *mm9520@fri.uni-lj.si, 63180206*

Introduction

Every online service benefits from predicting number of potential paying customers. Databox is no exception, which is why we were provided with data of their customers. In Project 2 we analysed customers' session behaviour. We have shown that on average number of *Sessions* starts to increase a few weeks before a customer becomes paying. Now we try to use this information with predictive modeling to see if we can predict when a customer becomes paying. We also extend this idea by not only looking at *Sessions* but also at *Calculations, Metrics, New Datasources, Queries, Urls Shared, Snapshots Shared* and *Users Added*.

We split this task into two parts. For the following month we try to predict number of new paying customers by:

1. computing the average of customers that become paying in the same month as they sign-up and
2. using predictive modeling with user data from previous months.

Methods

To predict the number of new paying customers in the next month we must take into account new customer sign-ups that become paying in the same month and customers that signed-up before and will now become paying.

Customers become paying in the same month

In the obtained Databox dataset there are **2,950** out of 101,698 customers which are paying. Customers usually need to become familiar with the online services Databox provides before becoming paying, this is why only **626** (**21%**) paying customers become paying in the same month as they sign-up. In Figure 1 we show a plot with number of new sign-ups, new paying customers and new paying customers that signed-up in that month and their average. We observe that all lines are nearly linear, which means that predicting new monthly sign-ups and paying customers who sign-up in the same month could be done by taking the average of previous months.

Customers from previous months

Since the majority of the paying customers become paying months after they sign-up our focus will be on them. We use their Databox site usage information to predict when will they become paying. We looked at customer's data 4 weeks before customer became paying. Our dataset contained customer's sessions information only from customers who signed-up after 20th April 2020. This is why we had to filter customers who signed-up before June 2020, since we need data from 4 weeks before they signed-up. We could not use other features like *had_trial* since we did not know the exact date they used their trial and we wanted to avoid using information from the future.

For each customer we constucted a boolean whether they became paying that month which was our target variable and number of specific event for each week for 4 weeks prior target month. We used data from June 2020 until May 2021 as training data and data from June 2021 until October 2021 as test data for each month separately. Since there is a big imbalance between paying and non-paying customers we used sampling. A lot of data for events is missing, so for non-paying customers we only used customers that had all of the

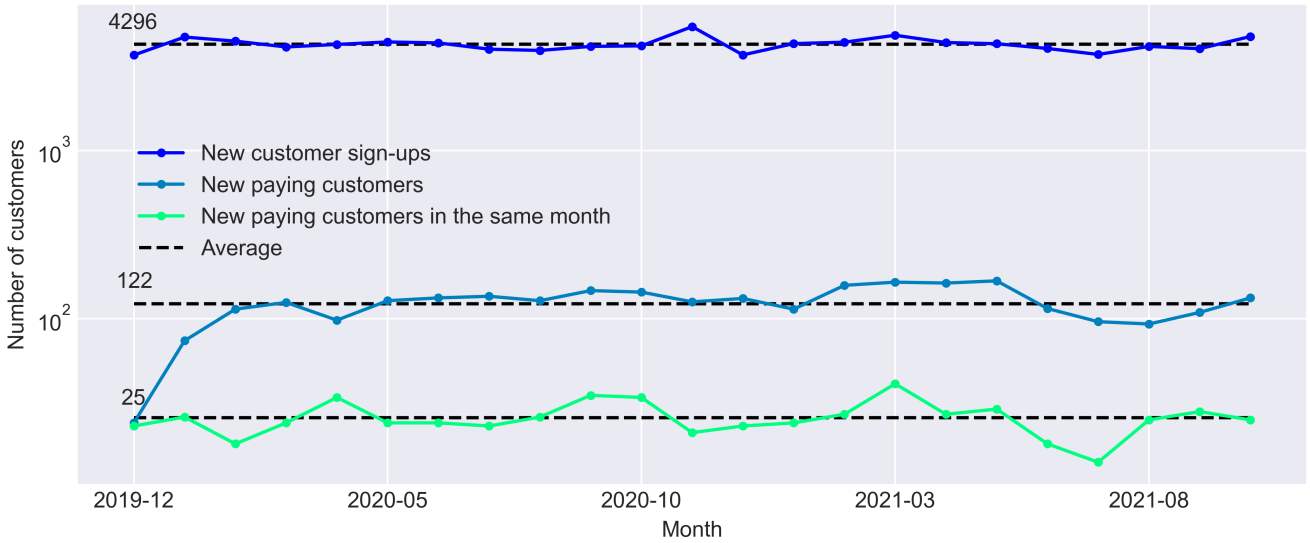

Figure 1. Number of new customer sign-ups, number of new paying customers and **number of new paying customer that have signed-up in the same month** for each month in the dataset and their average (dashed line).

events specified, which brought down the number to only **1991**. We also randomly chose the continous 4 weeks of events for non-paying customers and divided them into train and test set accordingly. Since we randomly sampled the 4 weeks and we wanted to get more robust results, we repeated training and predicting 10 times for each month.

Results

Customers become paying in the same month

We tried to predict the number of new monthly sign-ups by using new sign-ups and new paying customers from previous month, since we did not have any other data that could help us predict new sign-ups. Our hypothesis was that if the company was doing well previous month they would be doing even better the next month. With our KRR model we were able to surpass the baseline which was the average number of new sign-ups from training set.

Table 1. RMSE for prediction of number of new customers with Suport Vector Regression, Random Forest and Kernelized Ridge Regression with average from previous months as baseline.

Model	RMSE
SVR	425.5
RF	414.9
KRR	**383.2**
Baseline	404.0

As seen in Figure 2 the predictions from Suport vector regression and Random forest could not successfully predict the test set. Random forest captured the shape of new customer sign-ups but overestimated their value. In Table 1 we show RMSE for each model and our baseline. Since we do not have large enough dataset and RMSE between our models and baseline is not large, our deduction is that using the average number of sign-ups from train set is sufficient to predict new monthly sign-ups and new paying customers.

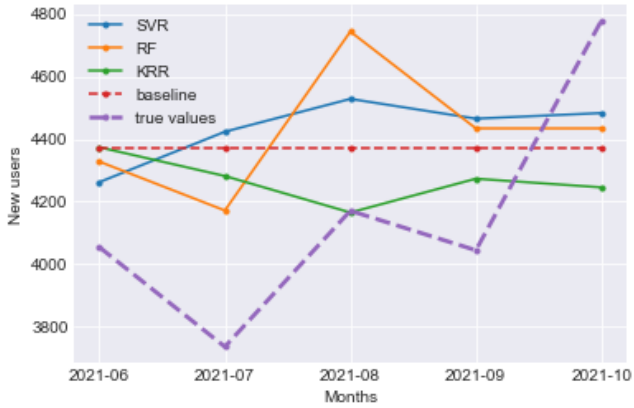

Figure 2. Prediction of number of new customers with Suport Vector Regression, Random Forest and Kernelized Ridge Regression with average from previous months as baseline and true values for each month.

Customers from previous months

If we want to predict number of paying customers for the following month the next step is to predict the number of new paying customers that have already signed-up. We used customer's event data to predict whether a customer will become paying. In Figure 3 we show that XGB and RF models were able to outperform the majority baseline classifier for all months except for August.

In Figure 4 we also show how close the prediction of

Figure 3. Accuracy with standard errors (transparency around lines) for Extreme Gradient Boosting, Random forest and Logistic Regression with majority baseline for predicting the number of new paying customers in the next month.

number of paying customers is for each model. From the plot it would seem that Logistic regression model learned how to predict the number of paying customers best out of models, despite having the lowest accuracy, which means that the other 2 models focused on predicting negative samples correctly.

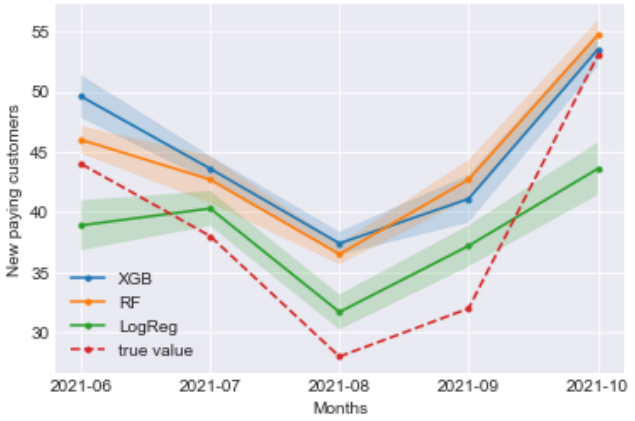

Figure 4. Prediction of number of paying customers with standard errors (transparency around lines) with Extreme Gradient Boosting, Random forest and Logistic Regression with true values for predicting the number of new paying customers in the next month.

In Figure 5 we show feature importance for first iteration of predicting for month June for XGB. We annotated all events by their initials and the number of weeks before becoming paying. We observe that the number of Queries that a company makes in the 4th week before becoming paying is the most important one, meanwhile number of Snapshots made has the least impact on predicting. We also observe that the 4th week in general is the most important out of 4 weeks. This could be the week a company starts getting more active which means they become paying in the next month.

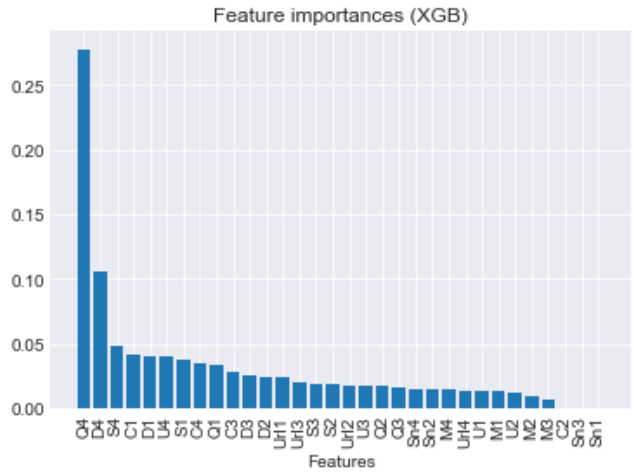

Figure 5. Feature importances for first iteration of predicting the number of new paying customers for month June for Extreme Gradient Boosting.

Conclusion

We focused on paying customers from Databox. We have shown that this prediction is not an easy task and then divided our work into 2 parts.

In the first part we tried to predict the number of new customer sign-ups and the number of new paying users that would sign-up in that month. We conclude that additional data would be needed for successful prediction and that using average from previous months is the best we can do with the data we have.

For the second part our task was to predict the number of new paying customers that have already signed-up. We used customer's event data and have successfully beaten the majority baseline, which proves that with a complete dataset we could successfully predict the number of new paying customers. Unfortunately the event dataset is severely lacking information from most users so even though we have beaten the baseline we could not push our models into production.

Our advice for Databox is that if they want meaningful results they need to provide better and more complete datasets. We were able to beat the baseline by approximately 5% with the scarce dataset provided, so imagine what we could do with a perfect dataset.