User churn forecasting for the company

Churn forecasting is one of the most important tasks in the field of working with audiences and relevant for the company

- The idea of this study is to learn to find customers which is planning to churn (If you learn to find such users with sufficient accuracy in advance, you can effectively manage churn);
- For finding such users we created the model which helps to forecast probability that the user will leave the service (this model is probabilistic binary classification model).

Description of the tasks and used metrics

We predicted the target class, which is the users leaving the service. The probability that the user belongs to the target class is the target value - the probability of churn. Accordingly, the greater this probability, the greater the chances that the user will refuse to use our service

For solving this tasks used a few metrics:

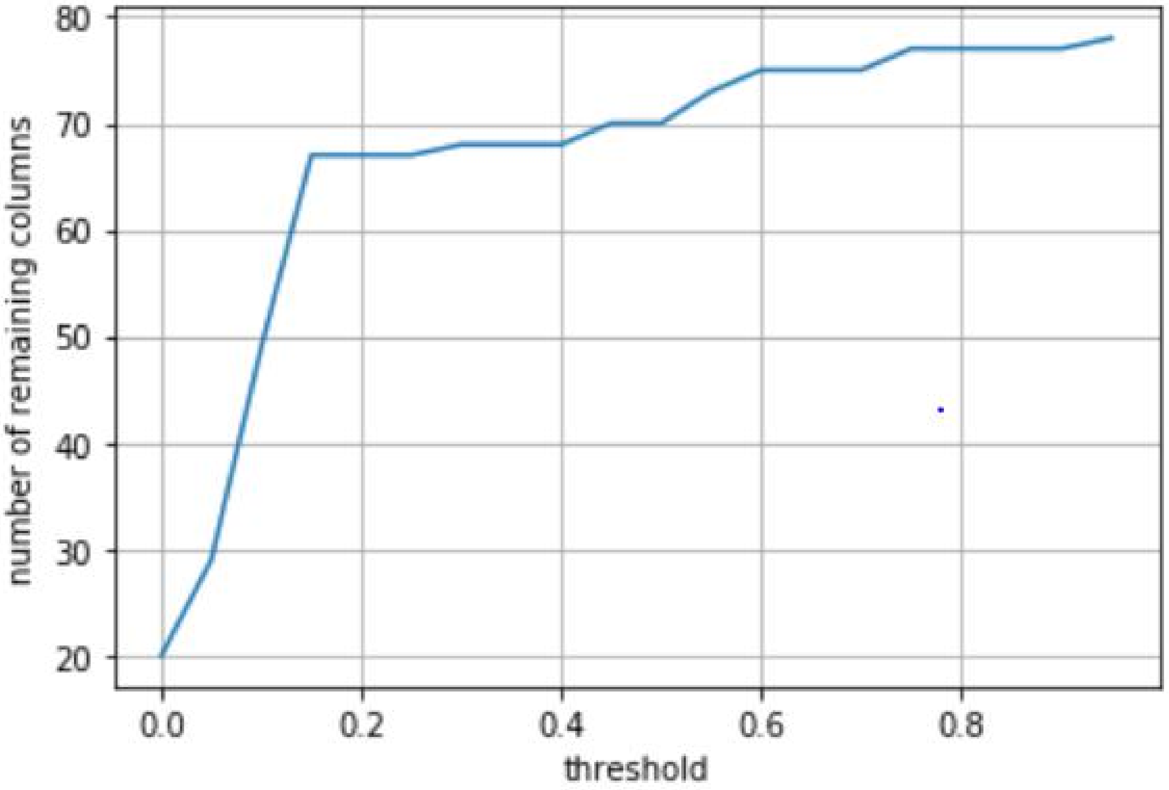
- ROC-AUC;
- Accuracy;
- F1.

But the ROC-AUC was the main metrics and we tuned models according to quality of ROC-AUC

Analysis and preparing the data

First of all, we should analyse the data. There are a lot of missings. We have to delete the columns which have missing more than selected threshold.

In the chart below presented how many columns will be deleted with different thresholds (75 % is the final decision).



Analysis and preparing the data

Secondly, we should analyse categorical features.

There are a few different approaches to encoding categorical features:

- One-hot encoding;
- Label encoding;
- Binary encoding.

After many attempts, it was decided to leave all categorical feaures and prepare that used one-hot encoding.

Analysis and preparing the data

Missings in the numeric columns were fiiled by means of every column.

In the table below presented the final performances with different approaches to fill missings.

ROC-AUC		
Mean	**Median**	**Zero**
0.735	0.7270	0.7348

Missings in the categorical features were filled by value '-100'.

Predictive modelling

There are of 3 different models to predict a churn:

- **Logistic regression;**
- **Random forest;**
- **XGBoost.**

In the table below presented the final quality of each models in hold-out datasets.

ROC-AUC		
Logistic regression	**Random forest**	**XGBoost**
0.6147	0.680	0.721

Predictive modelling

After analysing the candidates of the final model it was decided to use xgboost model. Finally, we should tuning the hyperparameters of the model using randomized search. In the table below presented the final hyperparameters of the final model.

subsample	min_child_weight	max_depth	gamma	colsample_bytree
0.8	10	4	0	0.8

ROC-AUC of the final model in hold-out dataset: **0.729***

* The final model has taken 9th place on the Kaggle competition

Ways to improve the quality of the model

There are a few approaches to improve the quality of the model.

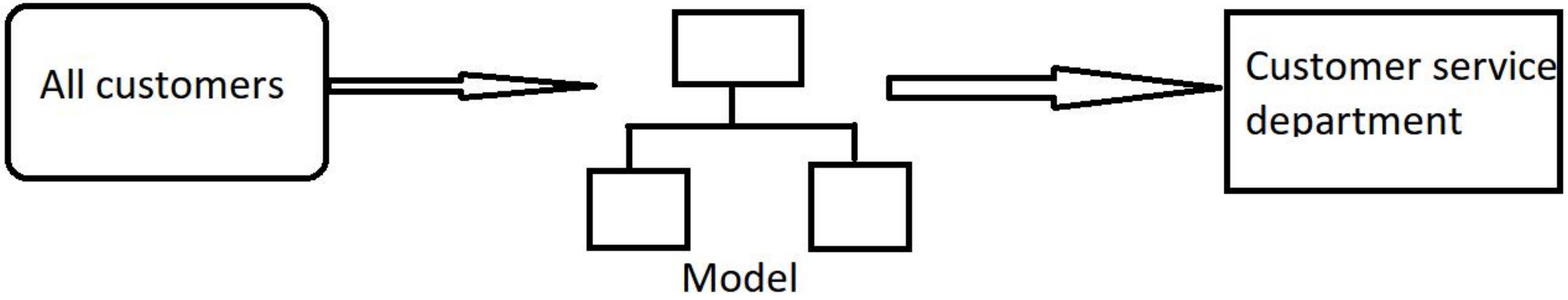
- Using the another realization of gradient boosting (e.x. catboost);
- Standartize the numeric columns;
- Encoding the categorical features using target based methods.

Application of the model

The final model will be used in the node 'Model”.

The process of using the model can be described as follows:

1. We will take all cutomers in company;
2. Then scoring process is happening; .
3. After it we should calibrate the result.s;
4. Then these results will be given to “ Customer service department”- they will analyze these scores and decide whether to make any suggestions to customersю



Conclusions

In this study, we analyzed the different characteristics that affect on user churn. The different approaches have been used to prepare the data for creating the model.

After conducting the A\B testing which can prove the effectiveness of our model it will be able to use in the company.

Economic effect. It is difficult at this stage to assess it. Everything needs to be tested on real users (it is necessary to perform A\B testing) and only then it will be possible to talk about the economic effect of implementation the model.