Test task (Data Scientist)
Yushkevich Natallia

CV:
https://drive.google.com/file/d/1qmeZXf1taQ8C_zPMA_upfHYp1zQlURjg/view?usp=sharing

Project:
https://github.com/YushkevichNV/A1-tasks

Task 1.

- Using historical data , build a time series model.
- Predict the daily behavior of the series in the next 3 months.
- Explain the choice of the forecasting method.
- Give an estimate of the forecast quality.

Task 1.

The data is displayed as a time series, where the x-axis is the day and the y-axis is the quantity.

Fig. 1.1 - Time series

Task 1.

The graph does not show any trend or seasonality (fig. 1.2). This suggests that the time series is stationary.

Fig. 1.2 - Components of the time series

Task 1.

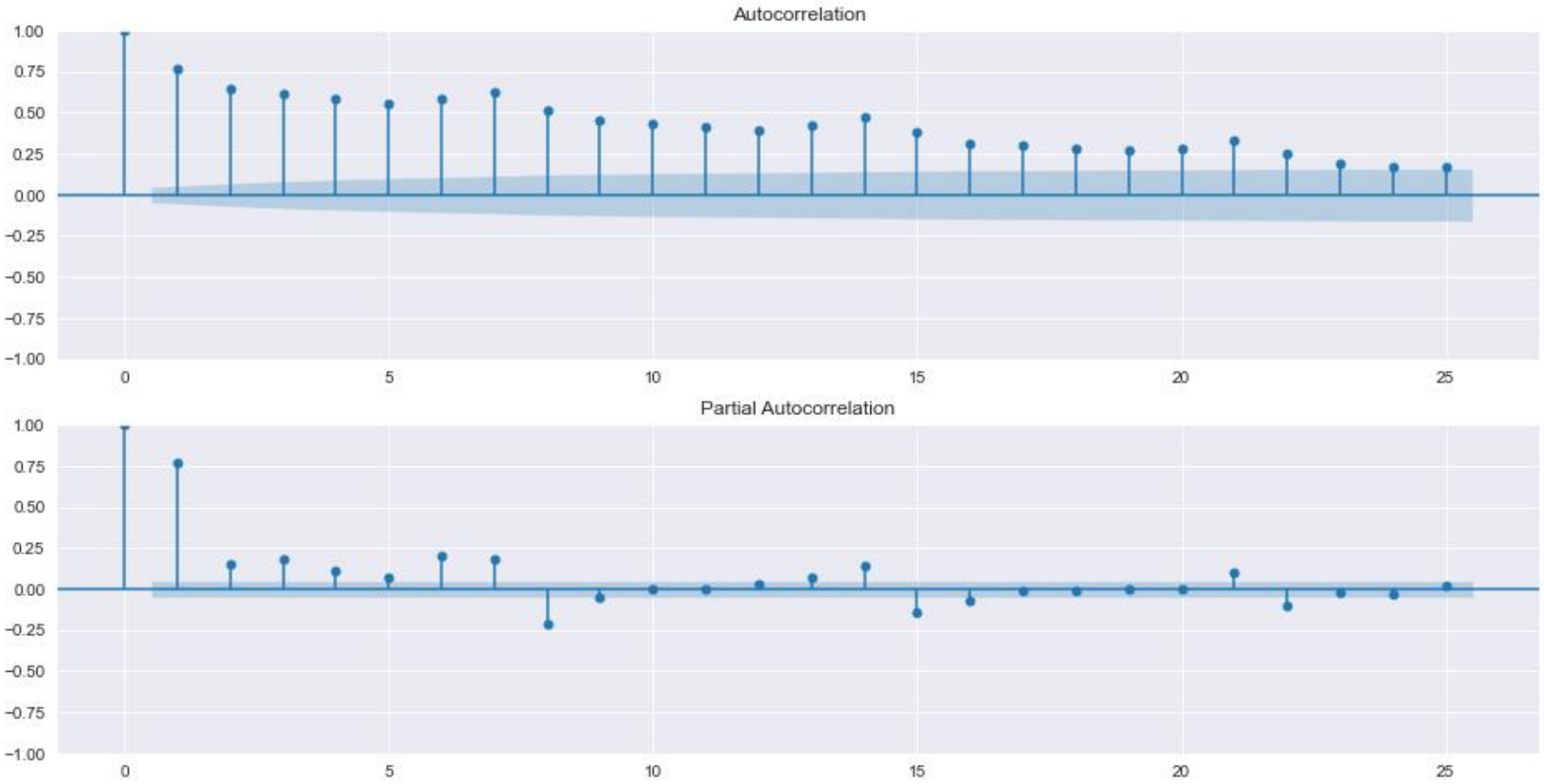


Fig. 1.3 - ACF, PACF

Task 1.

The Dickey-Fuller test showed that p-value = 0,000002.

p-value < 0.05, so the time series is stationary.

The PACF (fig. 1.3b) plot has a significant spike at lag 1 and less significant at p = [2; 8]. The ACF (fig. 1.3a) plot fades more smoothly. This may suggest an ARIMA(p, d, 0) model, where p can be taken from 1 to 8.

Since the series is stationary, d = 0.

Split the data into training and validation data:

- train_df = df[:'2019-03-31']
- test_df = df['2019-04-01':]

Task 1.

We trained the model ARIMA(8, 0, 0).

Metrics: MAPE = 12,77%, MAE = 414,17, MSE = 312250,57.

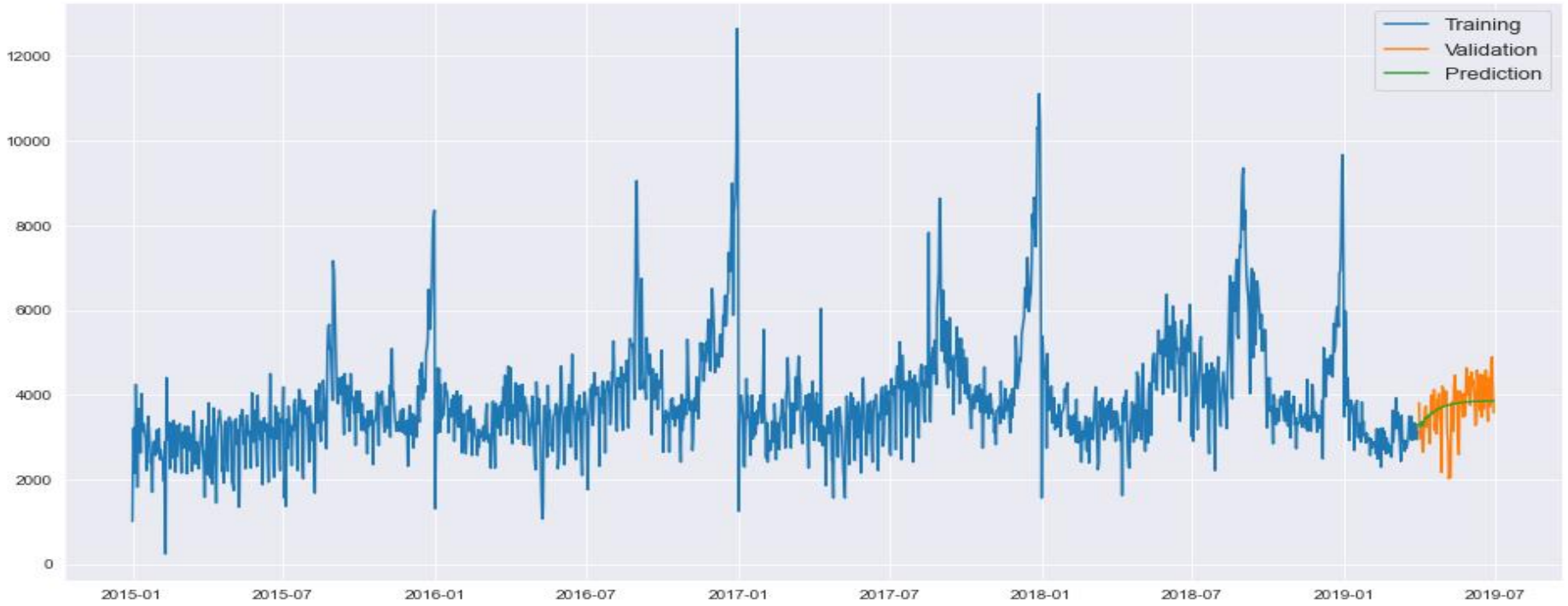


Fig. 1.4 - Training, valiudation and prediction

Task 1.

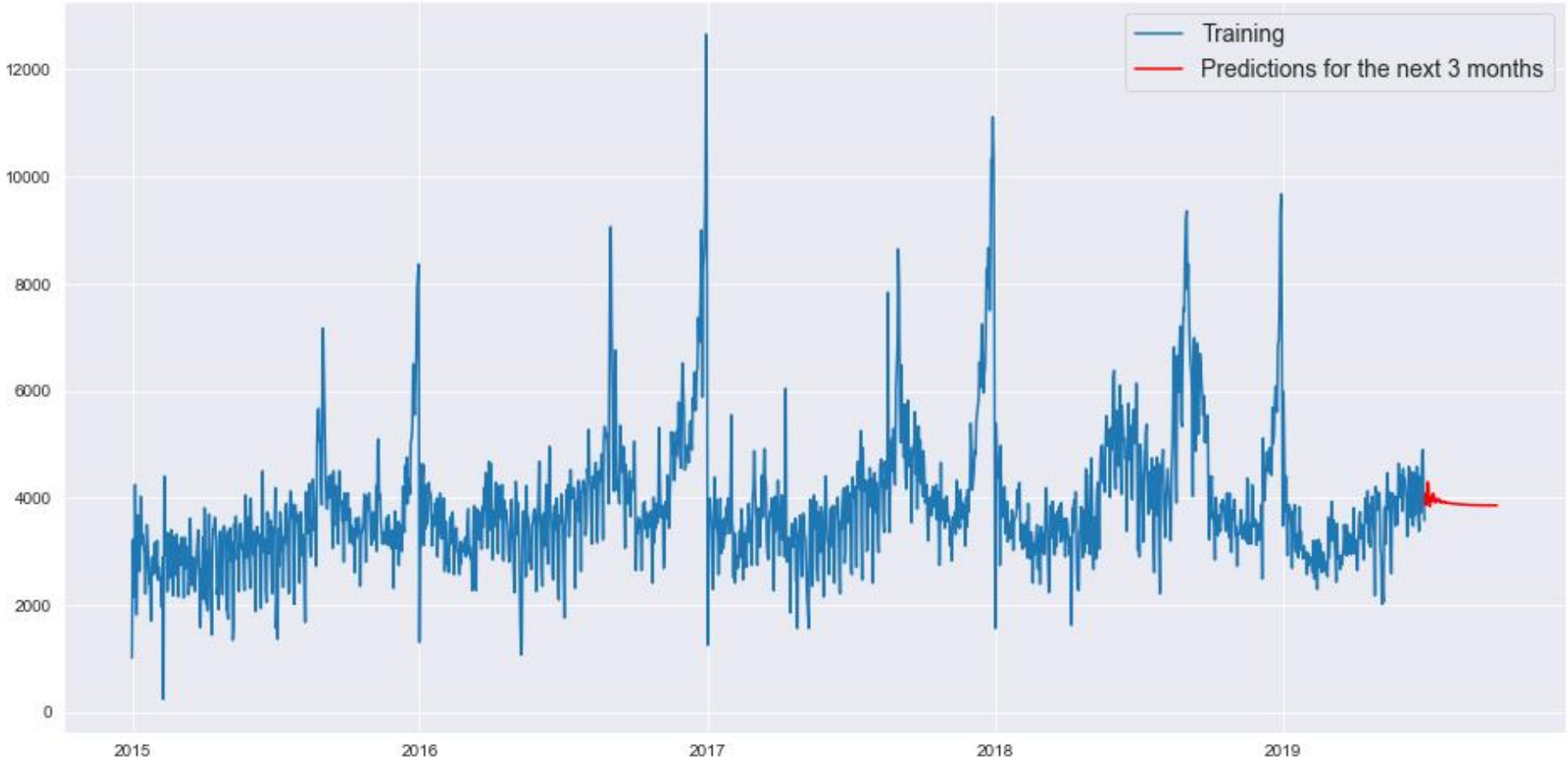


Fig. 1.5 - Finaly result

Task 2.

- Predict the values of the Target variable for the data set on the "Validate" sheet.
- Explain the choice of method.
- Give estimates of accuracy and quality of the predictive model.
- Construct an ROC-curve.
- Name the three most important predictors.

Task 2.

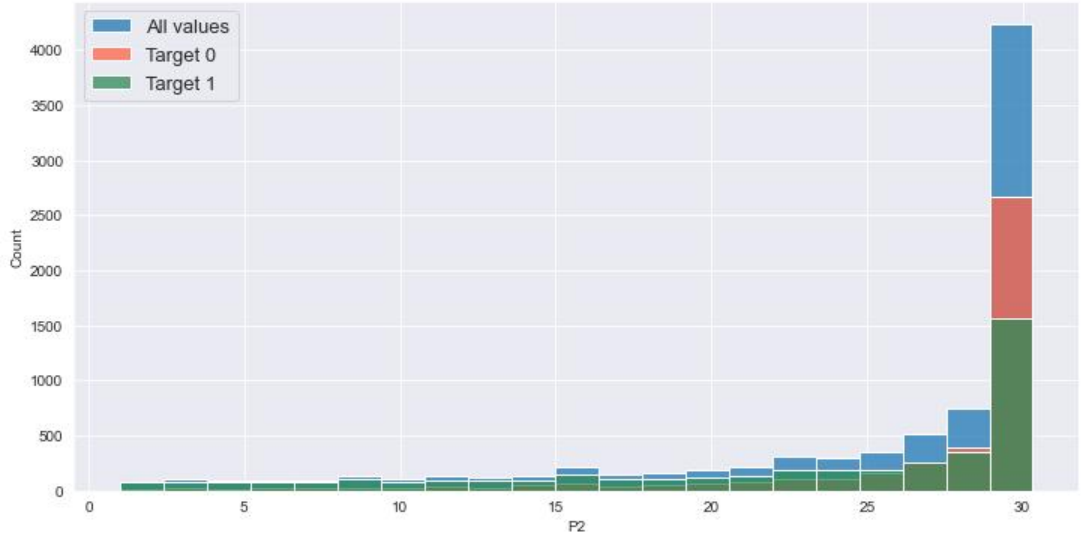
Preliminary analysis.

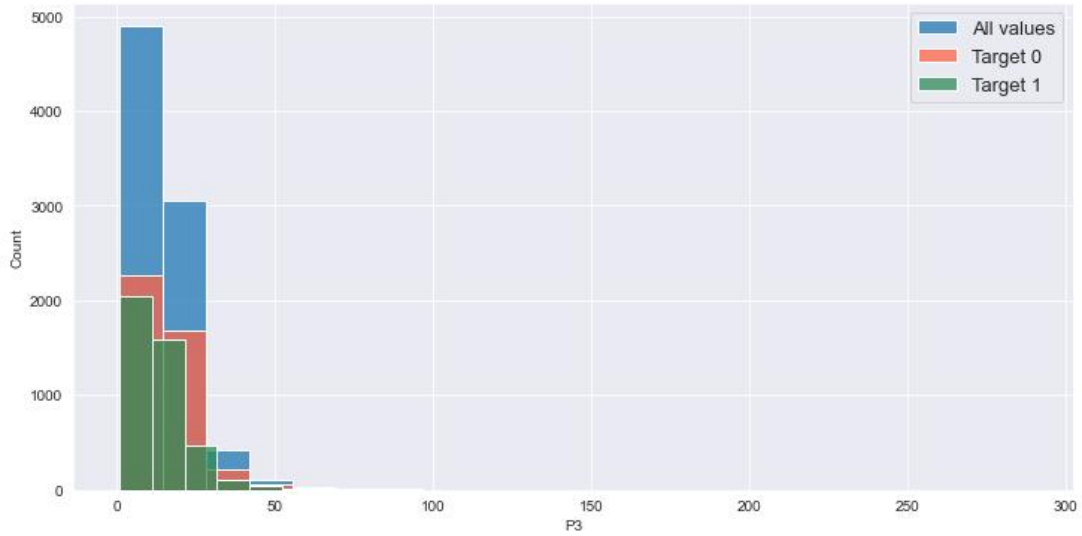
- Training datase has 10 000 rows.
- Validation dataset has 20 000 rows.
- No duplicate rows.
- All data is of type ‘float64’ or ‘int64’.
- No uninformative data.
- There is no accompanying documentation for the data, so the column values remain unclear.
- Target has values 0 or 1 (binary classification).
- There is no class imbalance.

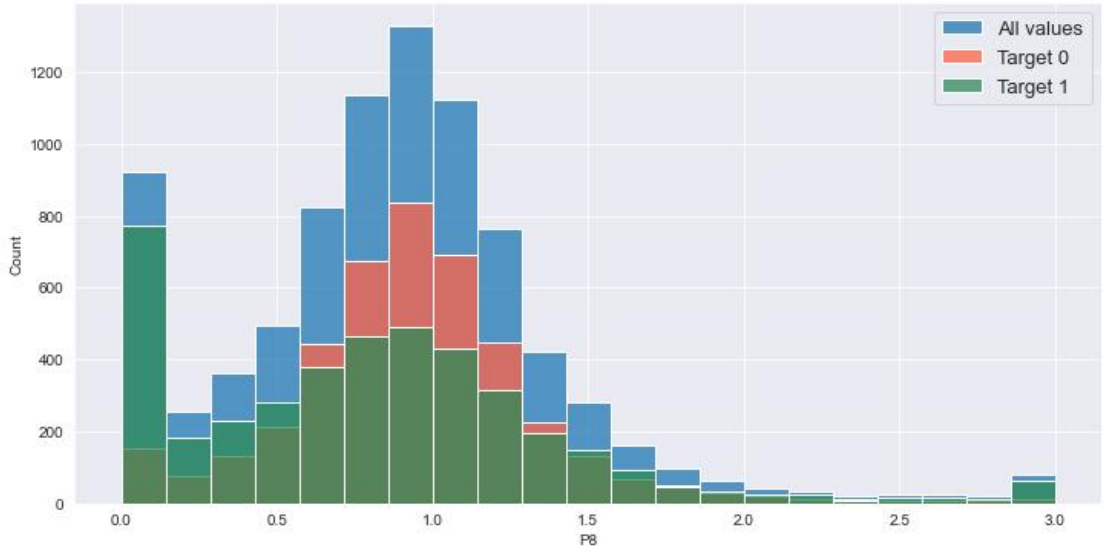
Task 2.

Missing values:

- P2 - 14.92 %
- P3 - 14.85 %
- P8 - 15.11 %



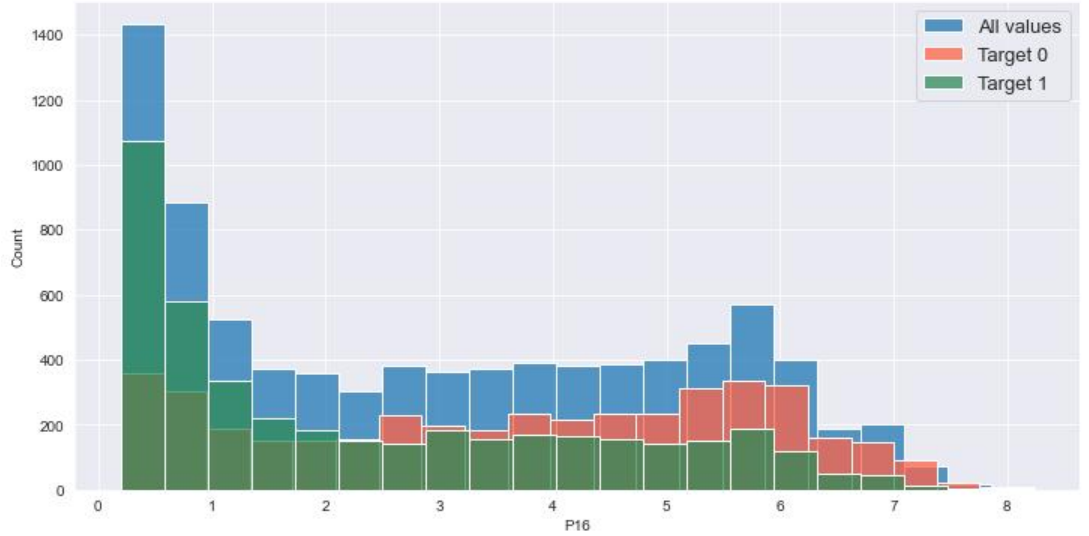


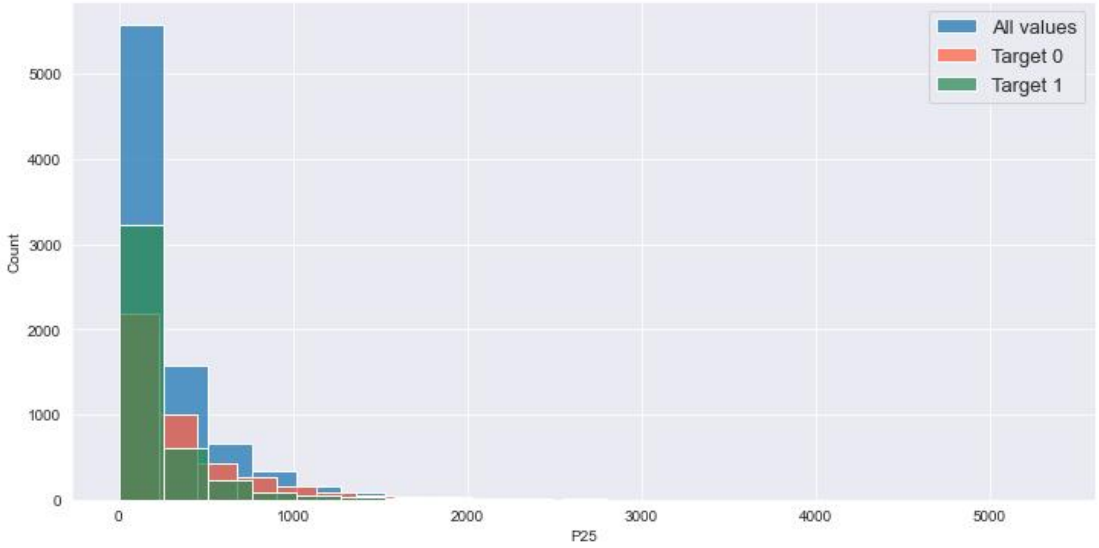


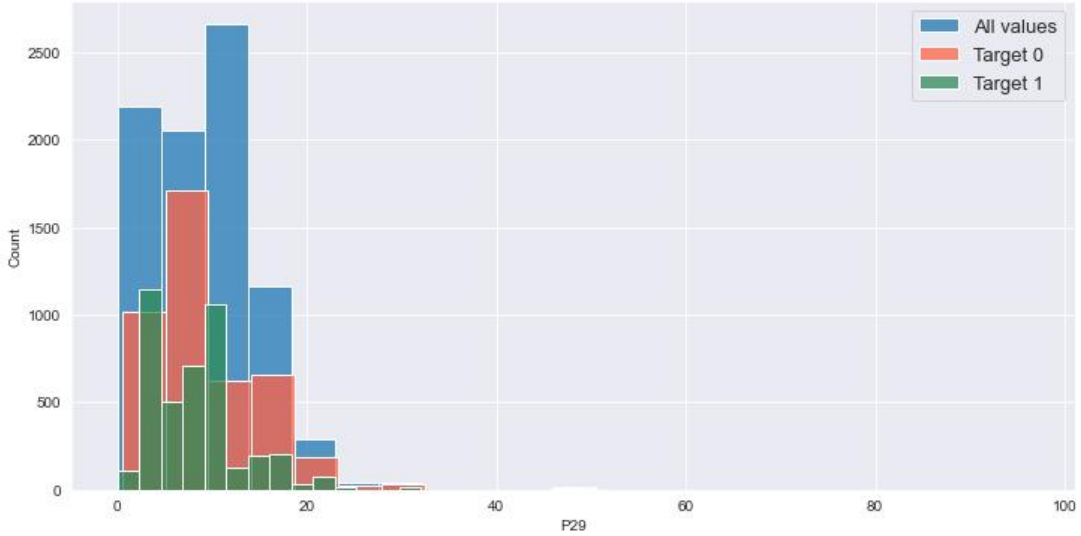
Task 2.

Missing values:

- P16 - 15.09 %
- P25 - 15.08 %
- P29 - 15.15 %







Task 2.

Missing values:

- We can say that the shapes of the histograms for the different target variables are similar.
- All distributions have offsets, so it is better not to replace the mean of the omissions.
- Replace the missing values with the median.

Task 2.

Correlating features:

- P1 → P5
- P17 → P25
- P22 → P23

- Delete the column P5, because P1 more informative.
- We leave the other columns in the dataset.

Task 2.

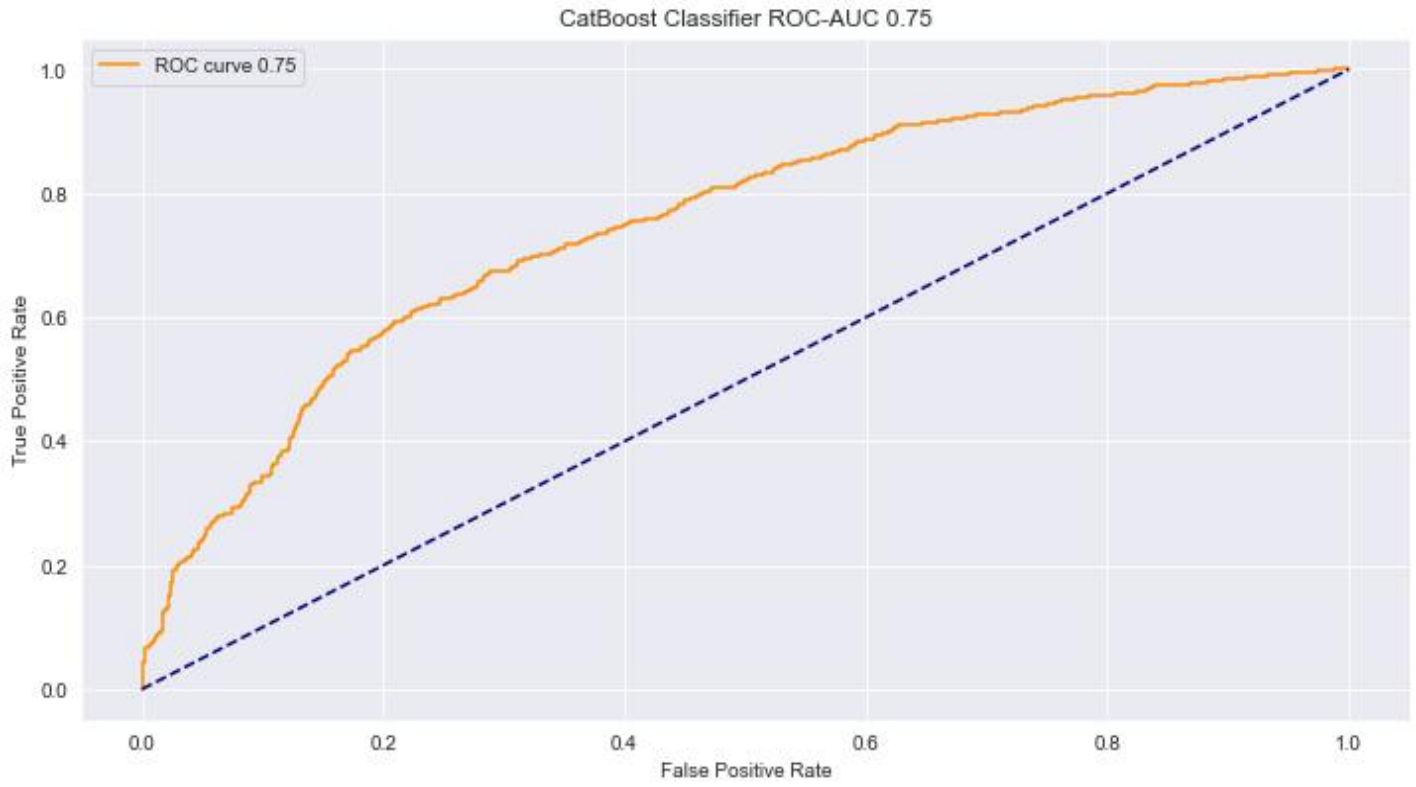
Model building:

- Use QuantileTransformer to make the data more even.
- We solve the problem of binary classification (classes are balanced).
- Binary classification approaches are usually similar to logistic regression models. We can also use the method of gradient descent.
- To evaluate the quality of the model we will use ROC-AUC (area under the error curve).

Task 2.

Baseline model CatBoostClassifier():

- accuracy = 0.688
- precision = 0.687
- recall = 0.692
- f1= 0.689
- roc-auc = 0.749



Task 2.

Selection of model parameters.

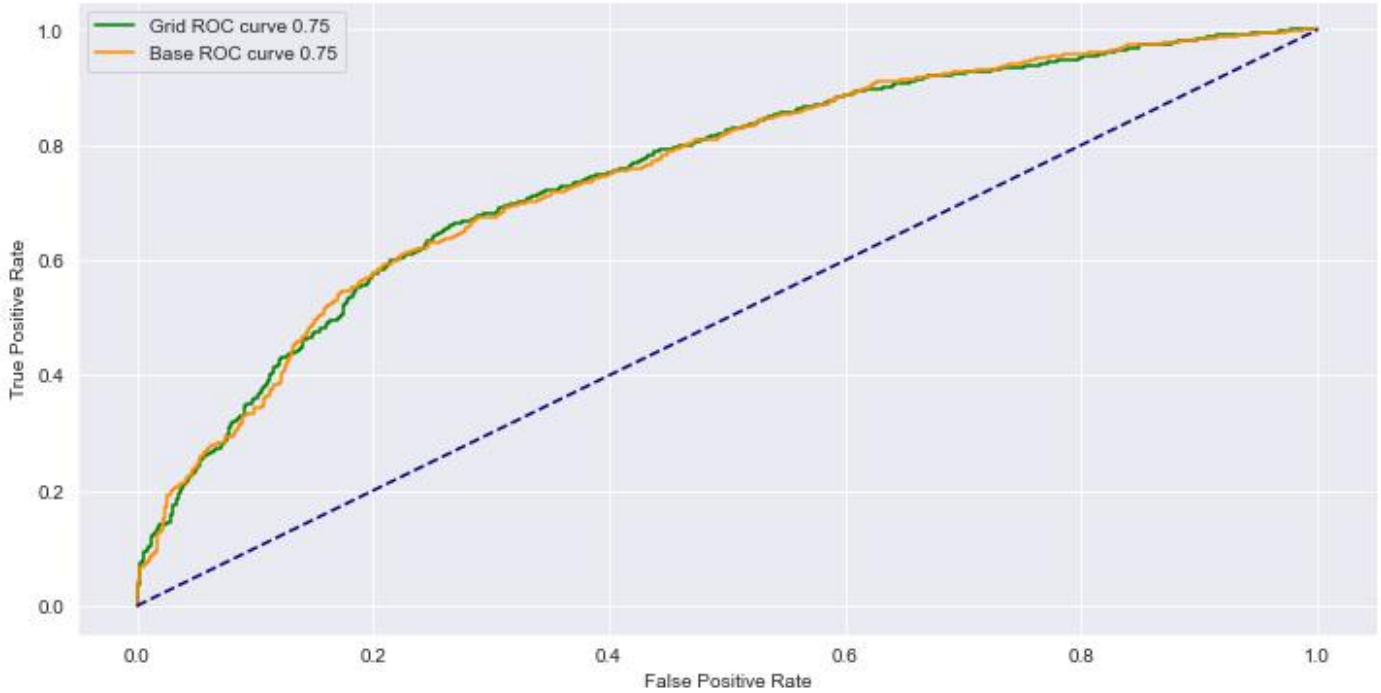
For the selection of model parameters we used GridSearchCV.

The best parameters:

- depth = 7,
- iterations = 1000,
- l2_leaf_reg = 2,
- learning_rate = 0.01

Task 2.

CatBoostClassifier(depth = 7, iterations = 1000, l2_leaf_reg = 2, learning_rate = 0.01):



- accuracy = 0.690
- precision = 0.686
- recall = 0.698
- f1= 0.692
- roc-auc = 0.750

All metrics have improved.

Task 2.

The three most important predictors.

We used ‘SHAP’ to find the most important features. The SHAP library calculates Shapley values to estimate the importance of a feature. To estimate the importance of a feature, the predictions of the model with and without the feature are evaluated.

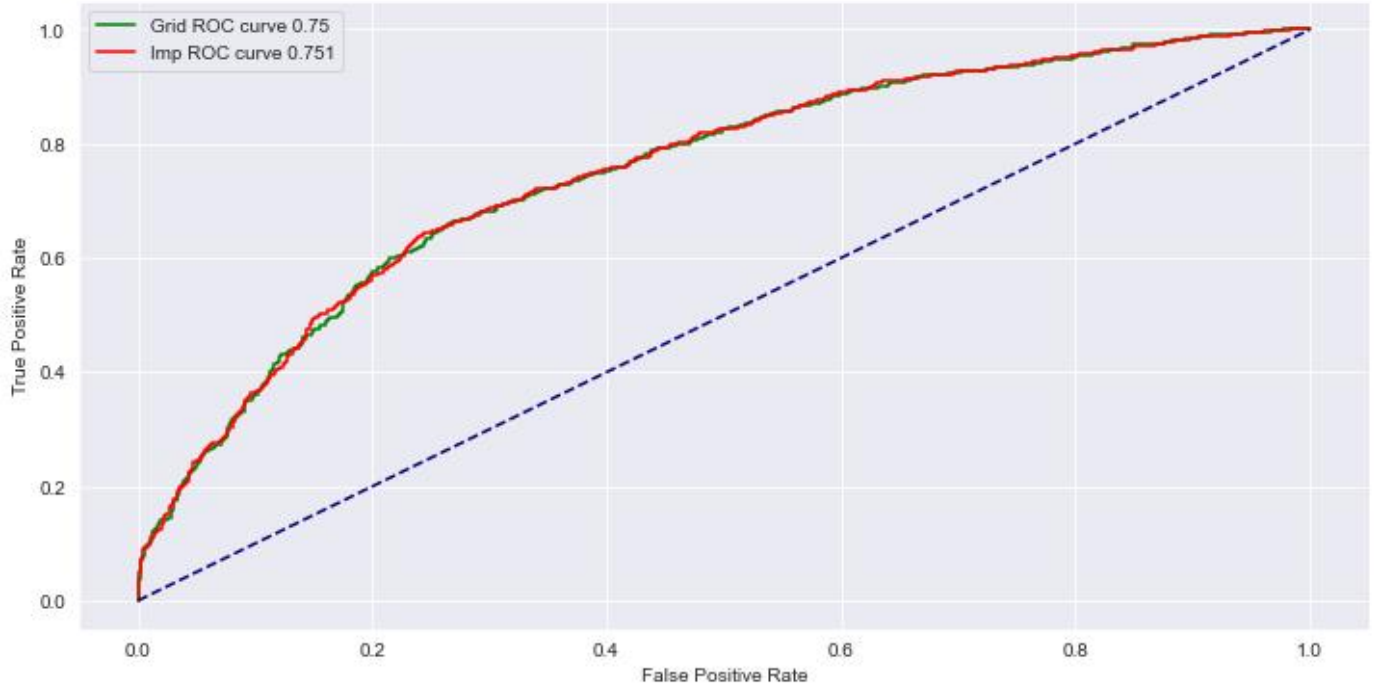
Top 3 predictors with a coefficient of significance:

- P16 - 0.314024
- P10 - 0.203953
- P23 - 0.187523

Task 2.

Let's train the model CatBoostClassifier(depth = 7, iterations = 1000, l2_leaf_reg = 2, learning_rate = 0.01) with features with a significance coefficient greater than 0.02.

- accuracy = 0.692
- precision = 0.692
- recall = 0.690
- f1= 0.691
- roc-auc = 0.751



All metrics (except f1) have improved.

Task 2.

Final model:

- CatBoostClassifier(depth = 7, iterations = 1000, l2_leaf_reg = 2, learning_rate = 0.01)

- Features with a significance coefficient greater than 0.02.

Task 3.

1. Directions for tariff plan changes.

Tariffs from which were switched and numbers of transitions:

- 1: 2059
- 2: 609
- 3: 2800
- 4: 662
- 5: 276

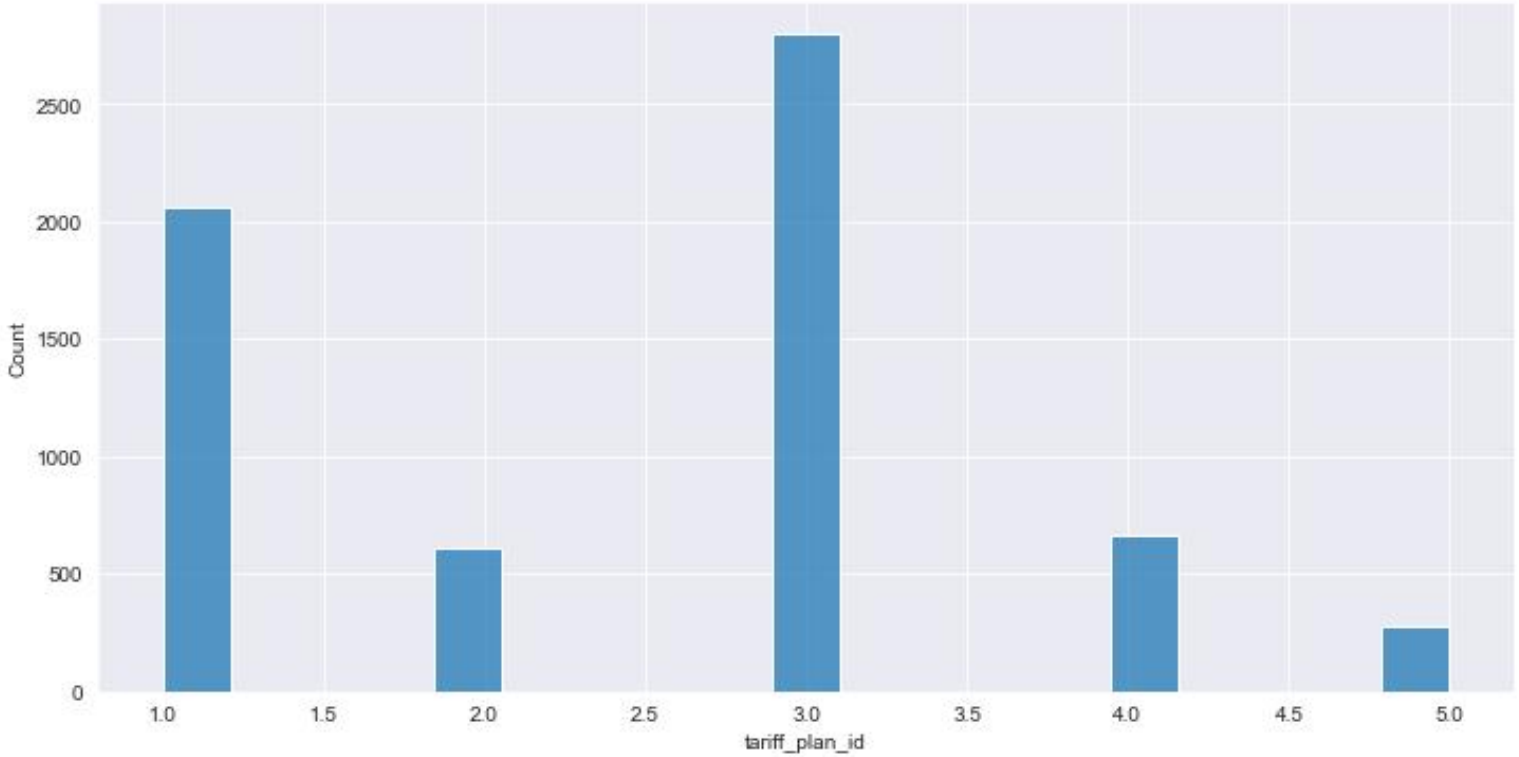


Fig. 3.1 - From which tariff plans there were transfers.

Task 3.

Tariffs to which the transition was madeand numbers of transitions:

- 1: 165
- 2: 41
- 3: 224
- 4: 958
- 5: 4952

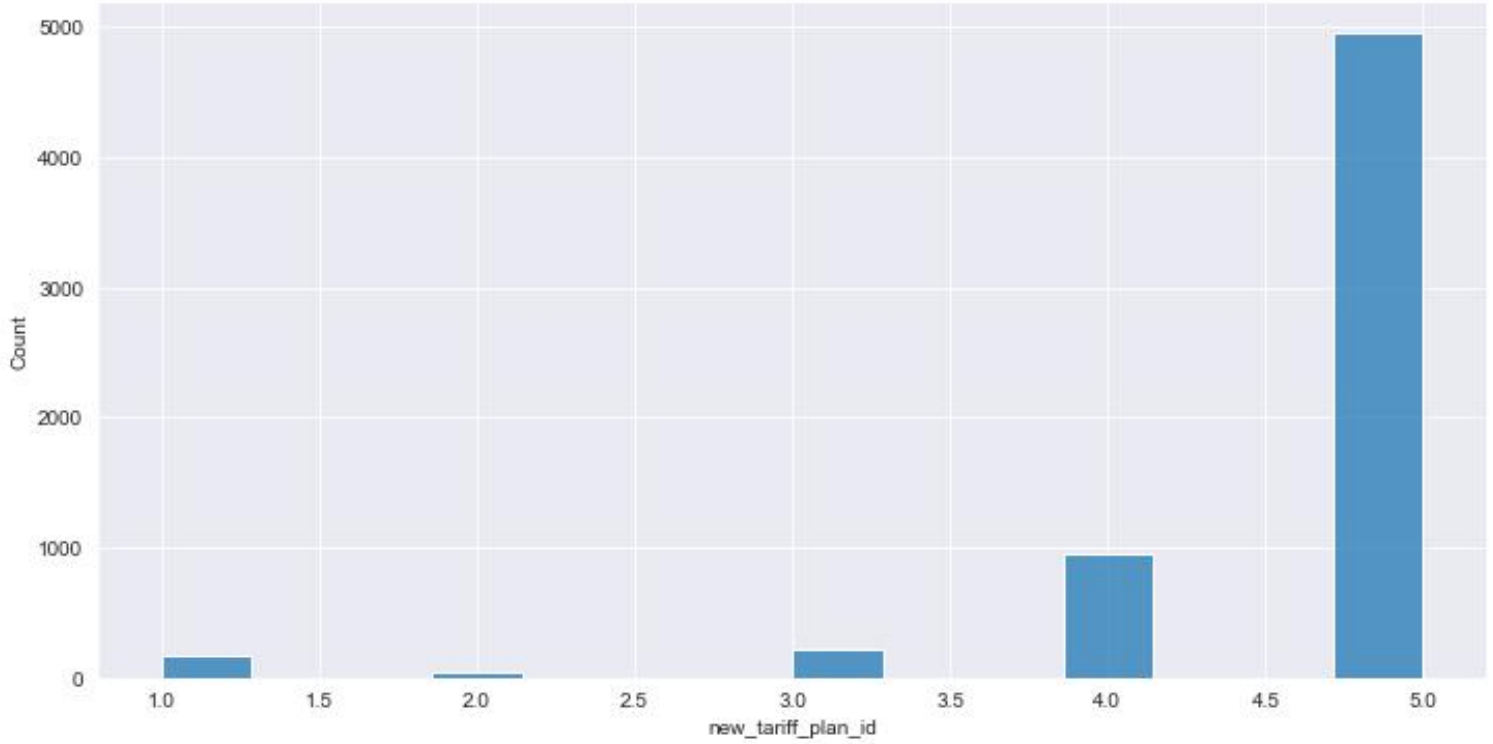


Fig. 3.2 - To which tariff plans the transition was made.

Task 3.

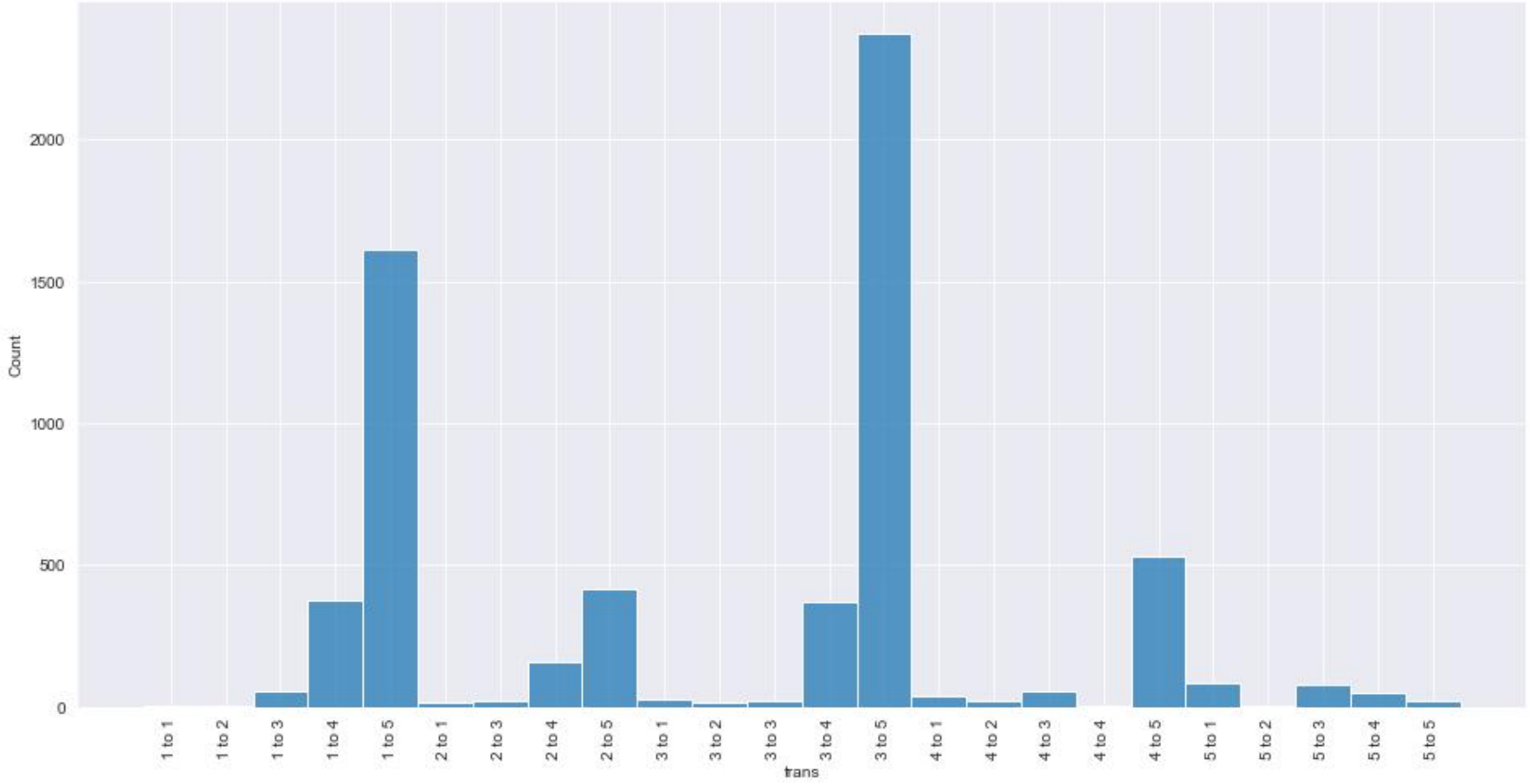


Fig. 3.3 - Transitions

Task 3.

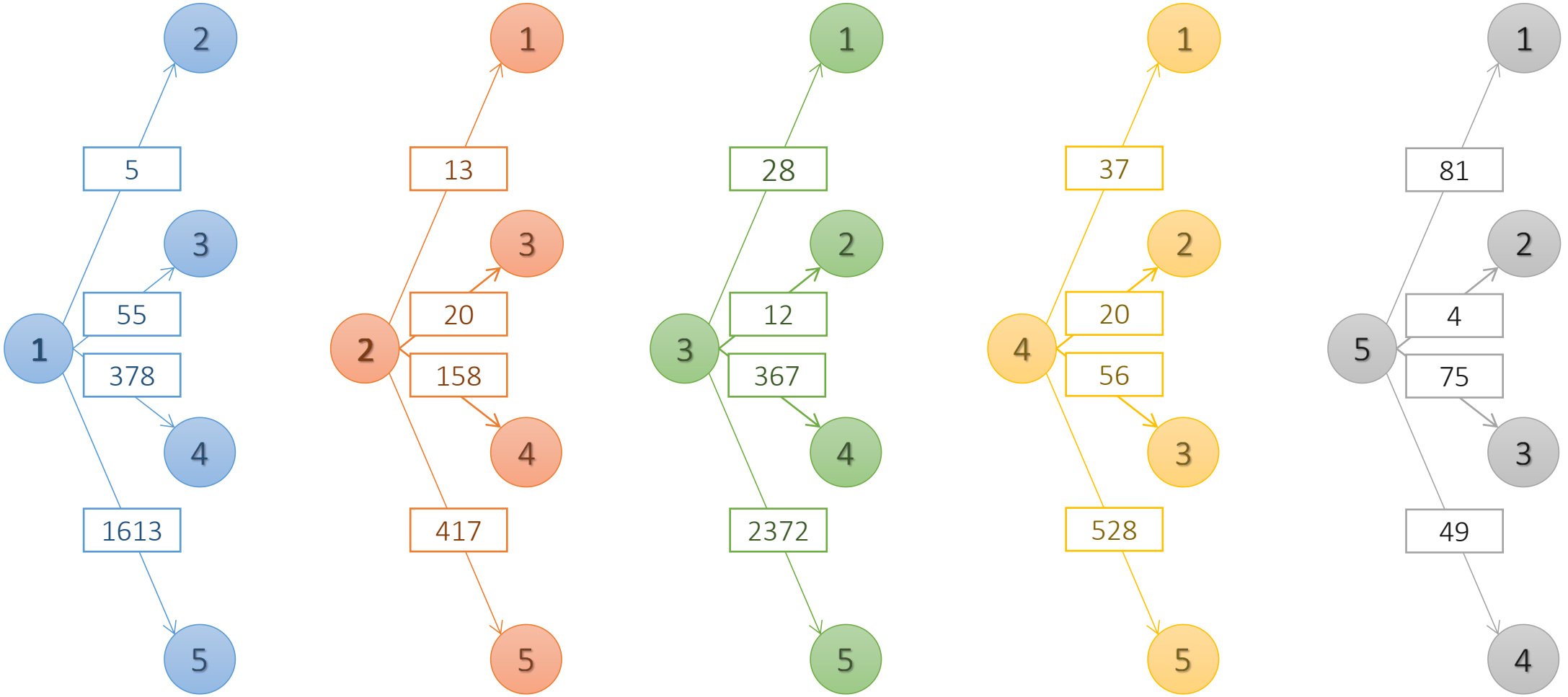


Fig. 3.4 - Number of transitions

Task 3.

- From the diagrams in fig. 3.3, 3.4 we can see that more often transitions were from tariffs with a smaller id to a tariff with a larger id.
- From this we can conclude that each successive plan is better than the previous one.
- The least number of conversions was from tariff 5 (fig. 3.2).
- From tariff 3 and 1 there were the most number of transitions (fig.3.1).
- There are transitions to the same plan (fig. 3.3). This indicates that the subscriber stopped using the services, but then came back.
- The most often transitions are: 3 $\longrightarrow$ 5 (2372), 1 $\longrightarrow$ 5 (1613), 4 $\longrightarrow$ 5 (528).

Task 3.

2. How much the average monthly bill of subscribers has changed?

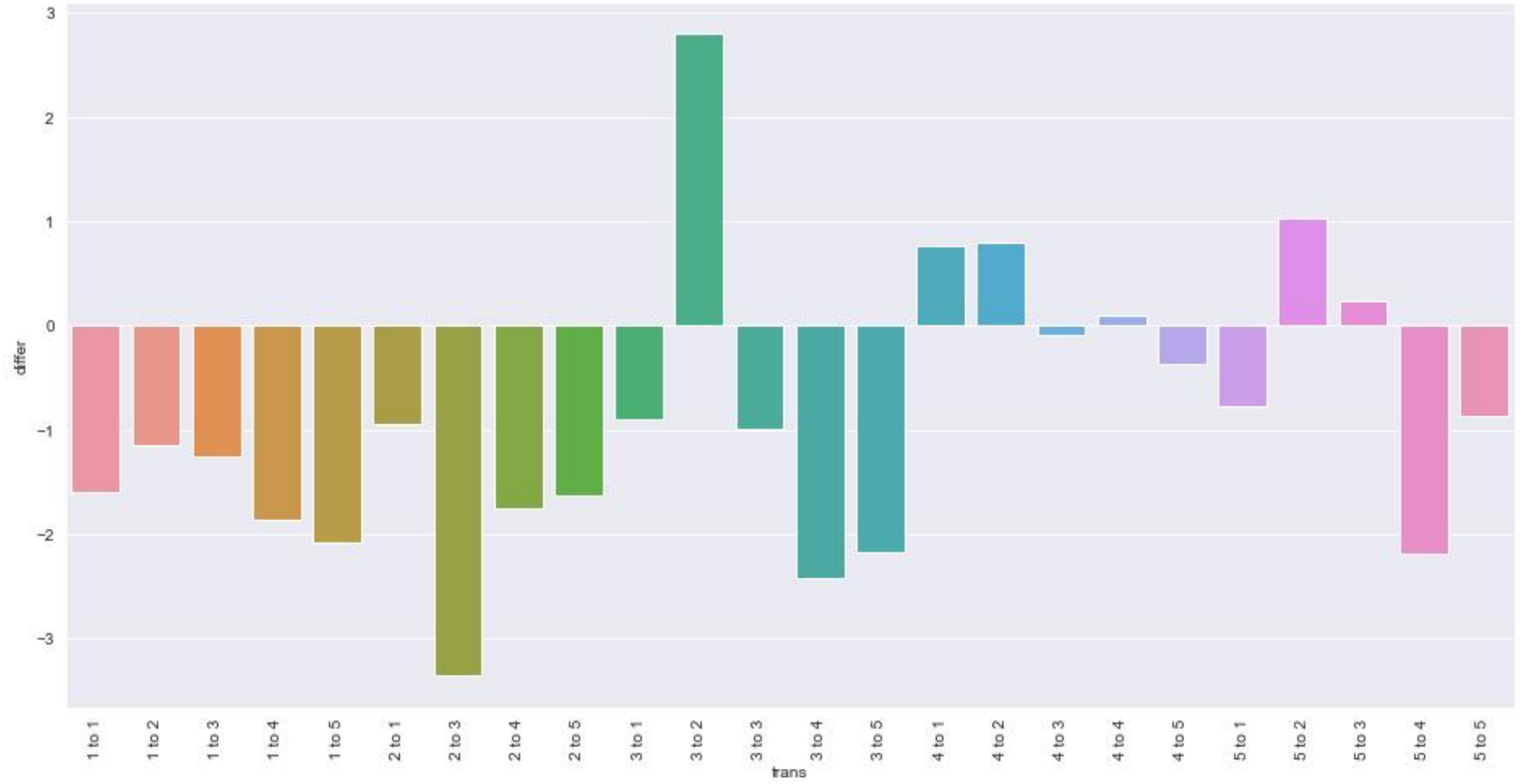


Fig. 3.5 - How much has changed in the average monthly bill of subscribers.

Task 3.

1 to 1	-1.591583
1 to 2	-1.139267
1 to 3	-1.260509
1 to 4	-1.852773
1 to 5	-2.070950
2 to 1	-0.946949
2 to 3	-3.348283
2 to 4	-1.753408
2 to 5	-1.634030
3 to 1	-0.898202
3 to 2	2.793472
3 to 3	-0.984714
3 to 4	-2.412072
3 to 5	-2.176663
4 to 1	0.760396
4 to 2	0.801367
4 to 3	-0.093776
4 to 4	0.093167
4 to 5	-0.366195
5 to 1	-0.775825
5 to 2	1.023000
5 to 3	0.231436
5 to 4	-2.191085
5 to 5	-0.869030

Table 1. Changes in average bill over a three-month period

Task 3.

- The directions of tariff plan changes, which are characterized by an increase in the average bill in the three-month period: $3 \longrightarrow 2$, $5 \longrightarrow 2$, $4 \longrightarrow 2$, $4 \longrightarrow 1$, $5 \longrightarrow 3$.
- The directions of tariff plan changes, which are characterized by an reduction in the average bill in the three-month period: $2 \longrightarrow 3$, $3 \longrightarrow 4$, $5 \longrightarrow 4$, $3 \longrightarrow 5$, $1 \longrightarrow 5$, $1 \longrightarrow 4$, $2 \longrightarrow 4$, $2 \longrightarrow 5$, $1 \longrightarrow 3$, $1 \longrightarrow 2$, $2 \longrightarrow 1$, $3 \longrightarrow 1$, $5 \longrightarrow 1$, $4 \longrightarrow 5$, $4 \longrightarrow 3$.
- The most expensive plan is 2.
- The most cheep plan is 4.

Task 3.

- Subscribers usually switched to a tariff plan with a smaller bill.
- Some people prefere to a more expensive tariff plan (3 → 2). Maybe the plan 2 includes more services then 3.
- From the fig. 3.3 we can see that a small number of people prefere more expensive tariff plan.
- Total bill change: -1.826

Task 3.

3. Changing the blocking frequency

Fig. 3.5 - Changing the blocking frequency

Task 3.

1 to 2	0.4
1 to 3	-0.01818182
1 to 4	-0.01587302
1 to 5	-0.02107874
2 to 1	0.30769231
2 to 3	0
2 to 4	0.15189873
2 to 5	0.03597122
3 to 1	0.10714286
3 to 2	-0.75

3 to 4	0.07084469
3 to 5	-0.03794266
4 to 1	0.05405405
4 to 2	0.05
4 to 3	0.07142857
4 to 5	0.04545455
5 to 1	0.07407407
5 to 2	0
5 to 3	0.10666667
5 to 4	0.10204082

Table 2. Changing the blocking frequency

Task 3.

- The number of blockages has increased: 1 → 2, 2 → 1, 2 → 4, 5 → 3, 5 → 4, 5 → 1, 4 → 1, 2 → 5
- The number of blockages has decreased: 3 → 2, 3 → 1, 4 → 3, 3 → 4, 4 → 2, 4 → 5, 3 → 5, 1 → 5, 1 → 3, 1 → 4.
- The number of blockages has not changed: 5 → 2, 2 → 3.
- General average change in the level of blockages -0.021.
- We can say that the number of blockages has not changed significantly.