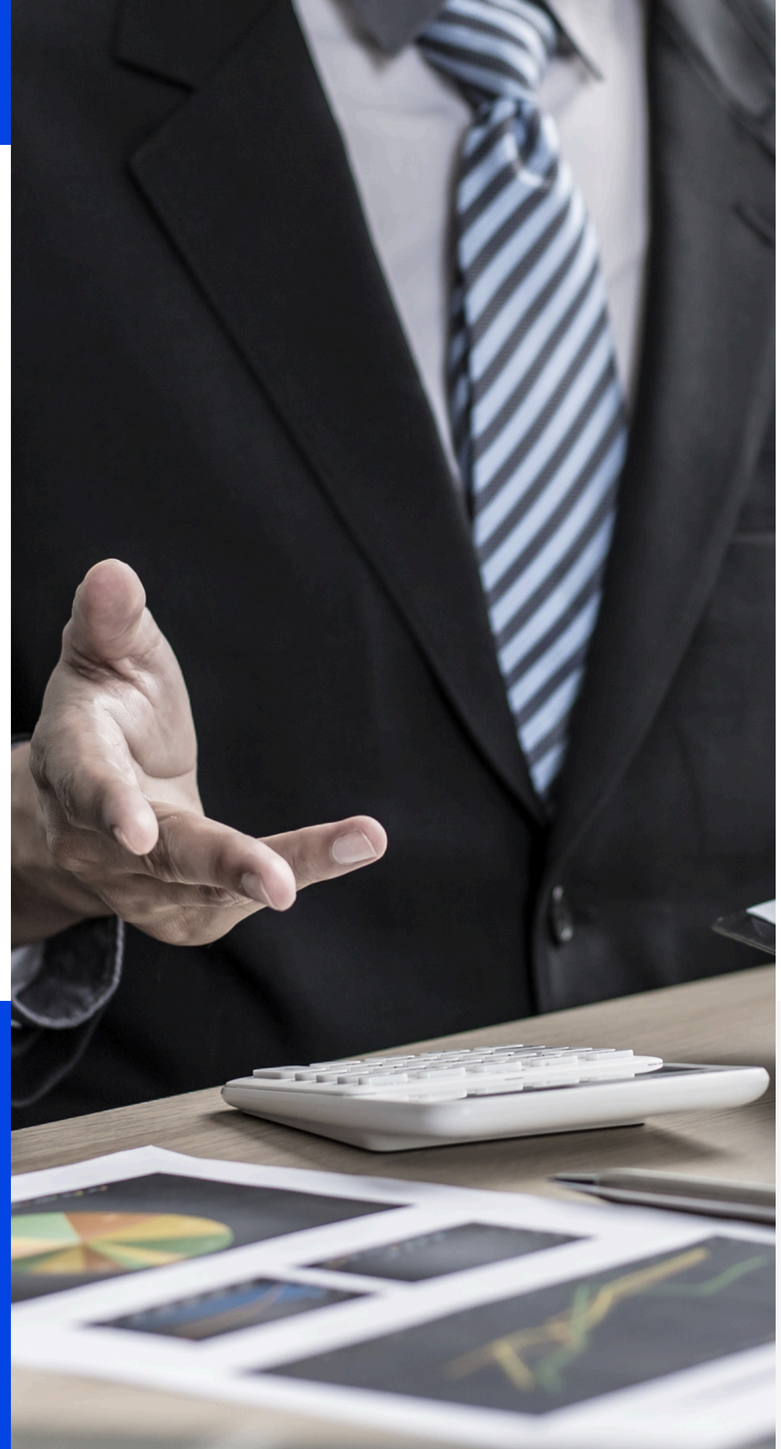
FINAL WORK

Umidjon Sattorov(student at Skillbox platform)

sattorov7474@gmail.com



CONTENT

BUSINESS PROBLEM

01 Hyphothesis testing

1) Organic traffic is no different from paid traiffic in terms of Conversion Rate(CR) to target events
2) Traffic from mobile devices is no different from traffic from desktop devices in terms of CR(Conversion Rate) into target events

02 Analytical questions

1)From which source/campaigns/devices/locations does the most targeted traffic come to us(both in terms of traffic volume and in terms of view CR)?

2)Which cars are in the greatest demand? Which cars have the best CR(Conversion Rate) for targeted events?

03 Machine learning model

Classification model for determining whether user in SberBankAuto website performed targeted actions like : 'sub_car_claim_click', sub_car_claim_submit_click', 'sub_open_dialog_click', sub_custom_question_submit_click' and so on.

DATA ANALYSIS

For the step data analysis and preparation, I have done following tasks.

Uploading

Initial analysis

Firstly, I have uploaded file to jupyter notebook and do some quick analysis work.(Checking the emptyness of columns and value distribution.)

Correlations

Decision-making for empty data

After quick initial analysis, I have found huge amount of empty values. Therefore I have accomplished some correlation analysis for making correct decision about the fate of these columns.

Cleaning

Filling missing values and cleaning data

After correlation analysis, I have dropped columns which has tremendous amount of empty values and removed the columns which doesn't have great correlation with targeted columns. Moreover, I have filled up missing values of some columns if possible.

Size reducing

Saving resources

After analysis and preparatory work in the dataset, I have reduced the number of unique values in some columns for saving computing resources.

ANALYSIS

DATA ANALYSIS AND PREPARATION

Uploading

```
The percentage of missing values in the dataset :

event_value                100.000000
device_model                99.220368
utm_keyword                 58.684721
hit_time                    58.400224
device_os                   58.388614
hit_referer                 39.753975
device_brand                25.156646
event_label                 23.765572
utm_adcontent               18.057880
utm_campaign                14.018759
utm_source                   0.004463
session_id                   0.000000
```

Correlations

	hit_type	hit_referer	hit_page_path	event_category	event_
hit_type	NaN	NaN	NaN	NaN	
hit_referer	NaN	1.00	0.80	0.48	
hit_page_path	NaN	0.80	1.00	0.54	
event_category	NaN	0.48	0.54	1.00	
event_action	NaN	0.32	0.35	1.00	
event_label	NaN	0.25	0.30	0.67	
visit_date	NaN	0.64	0.59	0.31	
visit_time	NaN	0.97	0.87	0.64	
utm_source	NaN	0.84	0.57	0.36	
utm_medium	NaN	0.71	0.60	0.31	
utm_campaign	NaN	0.52	0.55	0.37	
utm_adcontent	NaN	0.73	0.60	0.36	

Cleaning

```
The percentage of missing values in the dataset :

session_id                  0.0
hit_date                    0.0
geo_city                    0.0
geo_country                 0.0
device_browser              0.0
device_screen_resolution    0.0
device_os                   0.0
device_category             0.0
utm_adcontent               0.0
```

Size reducing

Unique values table

Column	Initial unique values	Last unique values
hit_page_path	20348	200
visit_time	38654	3

ANSWERING DATA ANALYSIS QUESTIONS

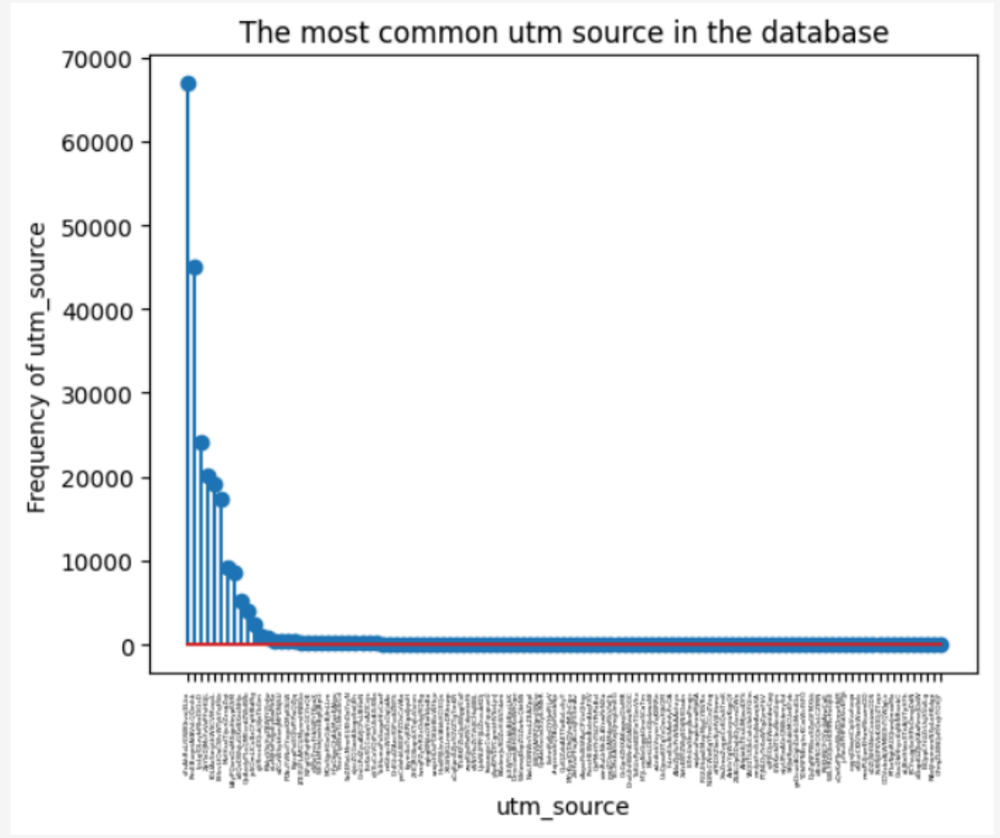
Hyphothesis testing

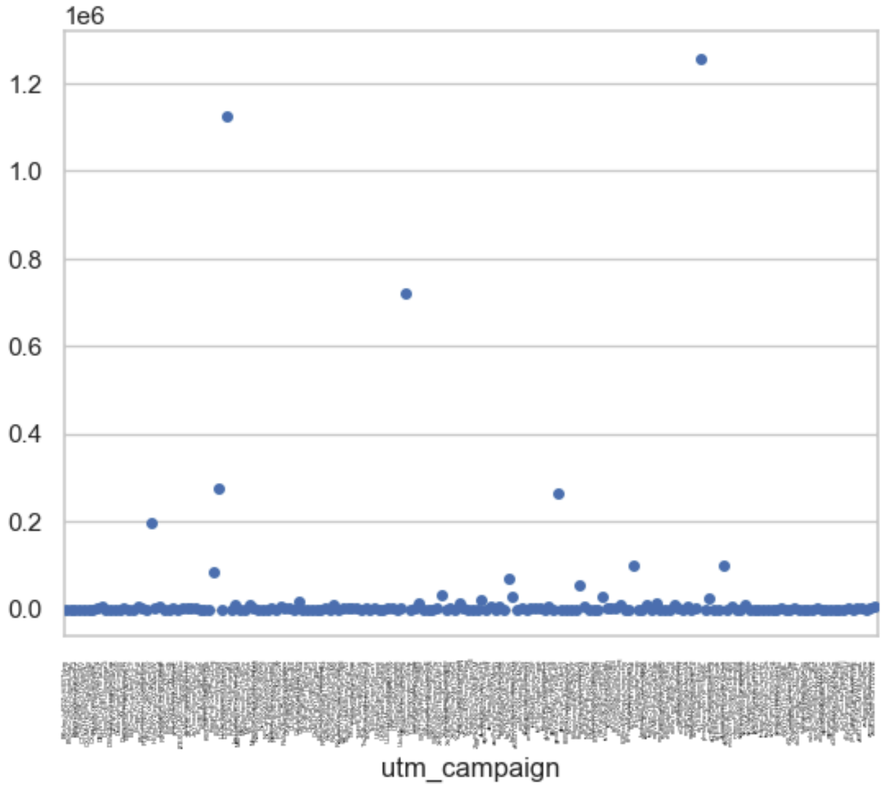
NULL hyphothesis	Alternative hyphothesis	p_value(Cramer's V algorithm)	True hyphotheis
Organic traffic is no different from paid traffic in terms of Conversion Rate(CR) to target events.	Organic traffic is distinct from paid traffic in terms of Conversion Rate(CR) to target events.	0.0	Organic traffic is not distinguishable from paid traffic in terms of Conversion Rate(CR) to target events.
Traffic from mobile devices is not distinguishable from thee traffic from desktop devices.	Traffic from mobile devices is distinguishable from thee traffic from desktop devices.	1.6e-07	Traffic from mobile devices is distinct from the traffic from desktop devices.
Traffic from Moscow and St. Petersburg is not different from each other.	Traffic from Moscow and St. Petersburg is different from each other	1.07e-21	"Traffic from Moscow and St. Petersburg is different from each other"

Hyphothesis analysis

In order to test the hyphothesis for probability, I have used Cramer's V algorithm.

ANSWERING DATA ANALYSIS QUESTIONS





Most common utm_source

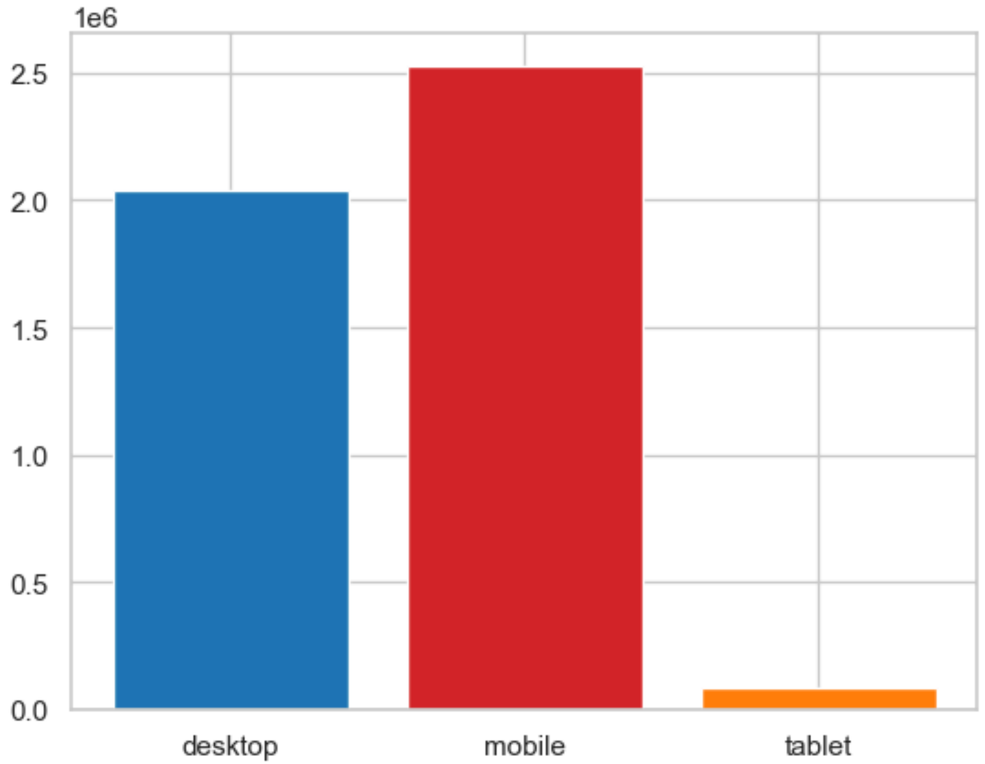
From the little stem chart, we can easily infer that the prevalent source of utm which the traffic come from is

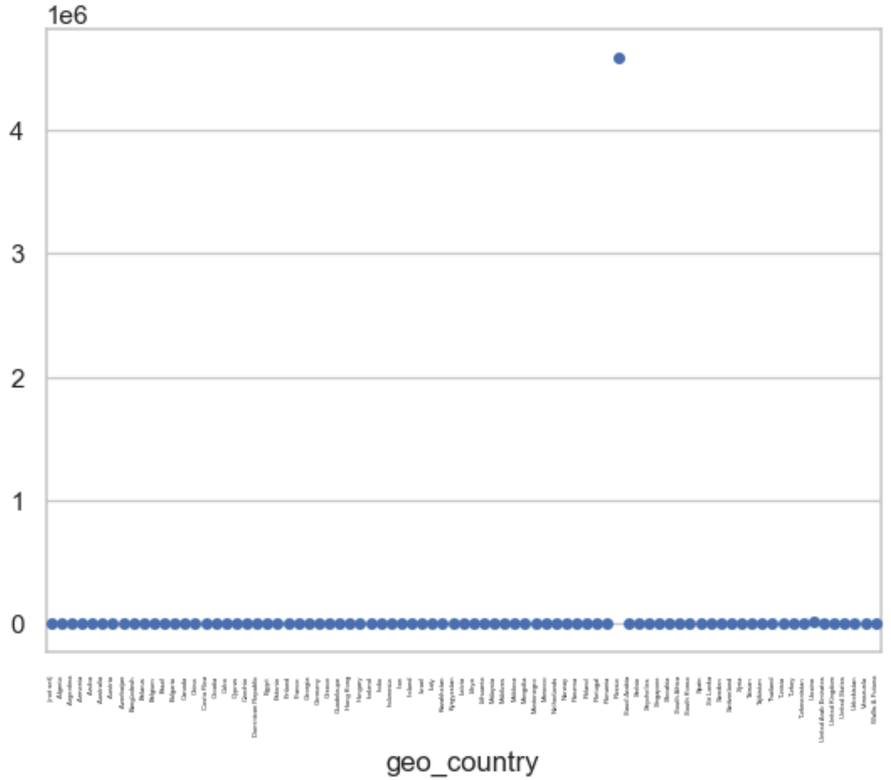
vFcAhRxLfOWKhvxjELkx

Most common utm_campaign

From the given scatter plot, we can easily conclude that the most traffic to the website is come from **okTXSMadDkjvntEHzIjp**

ANSWERING DATA ANALYSIS QUESTIONS





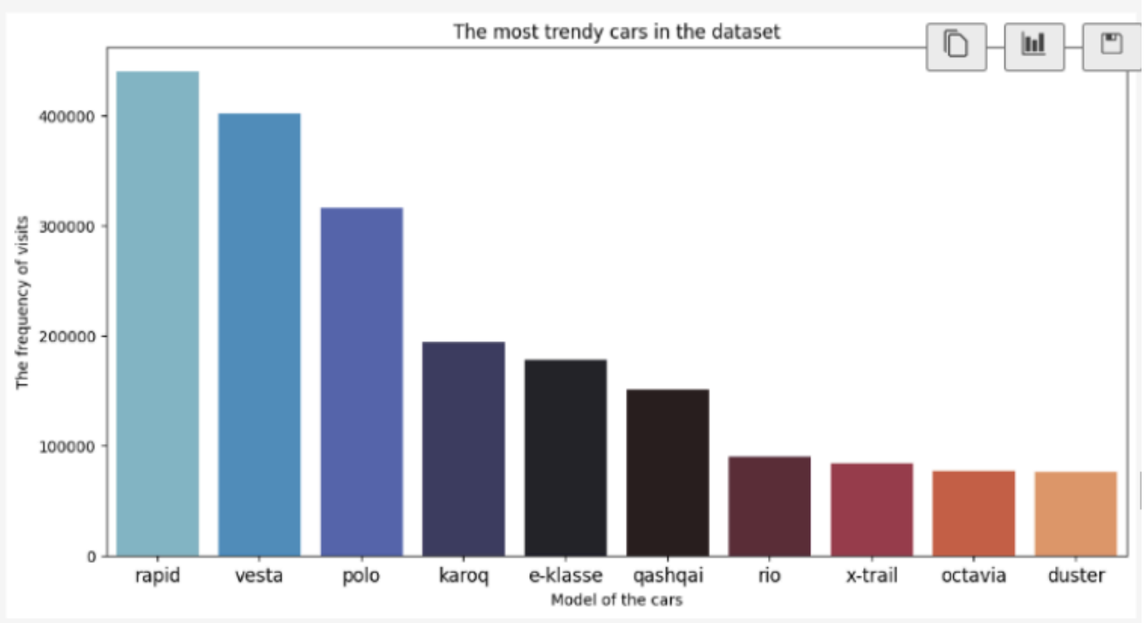
Most common device

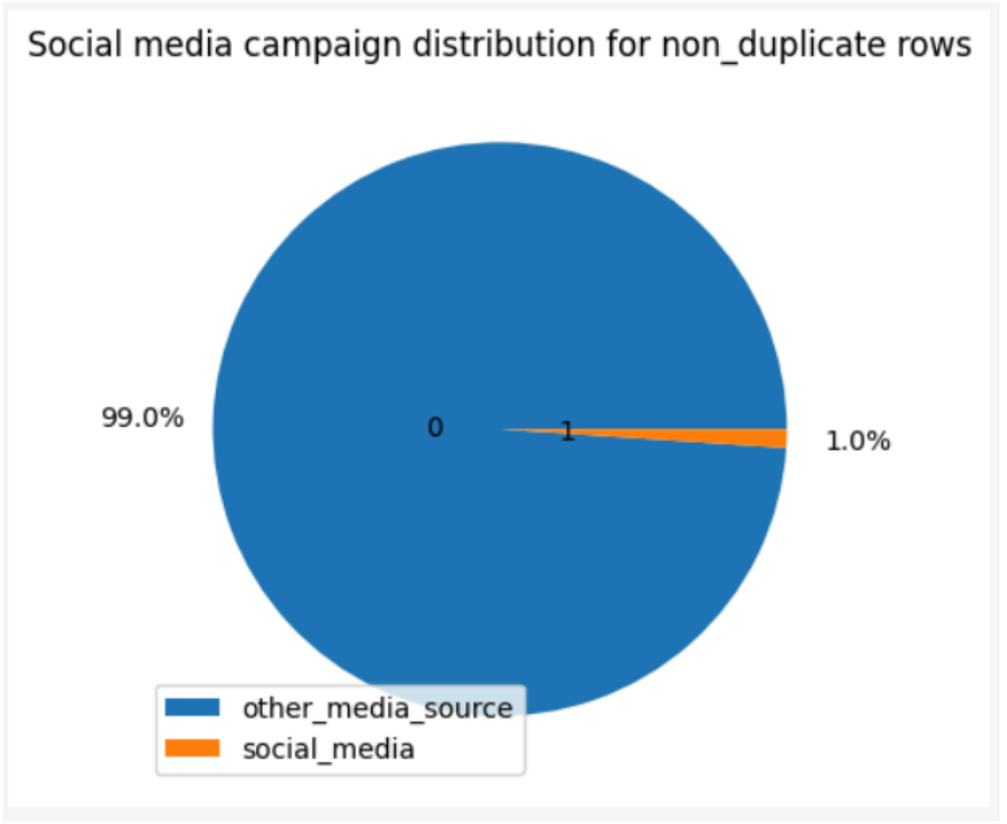
Given bar chart represents that **mobile devices** were the dominant category of technology which users used when accessing the website.

Most common country

Scatter plot illustrates that the most people to the car renting website were the **Russian** residents.

ANSWERING DATA ANALYSIS QUESTIONS





The most demanded cars

Bar chart indicates that the most trendy model of cars which was rented by website users were **rapid** and **vesta**.

Do the company need to invest to social media campaign?

As the contribution of social media to overall traffic to the website doesn't comprise huge percentage, it is better to invest to other marketing campaigns.

FEATURE ENGINEERING

In these stage of modelling, I have transformed the dataset by reducing the number of excessive columns and then do some feature engineering tasks.

One Hot encoder:

For this stage, I have replaced the values of table data to binary form using One Hot encoder. But before that, In order to avoid using extra columns, I have diminished the number of cols.

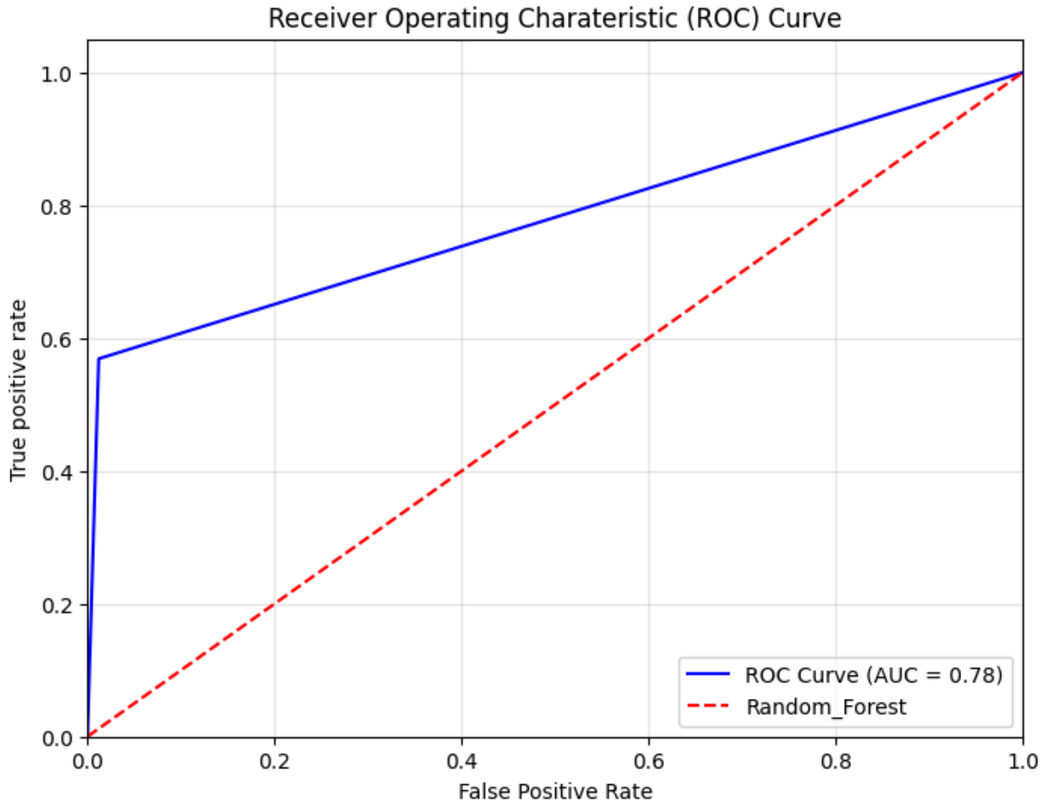
Standart scaler:

After using One hot encoder to categorical values, I have used Standart scaler for the values of numerical columns.

Label encoder:

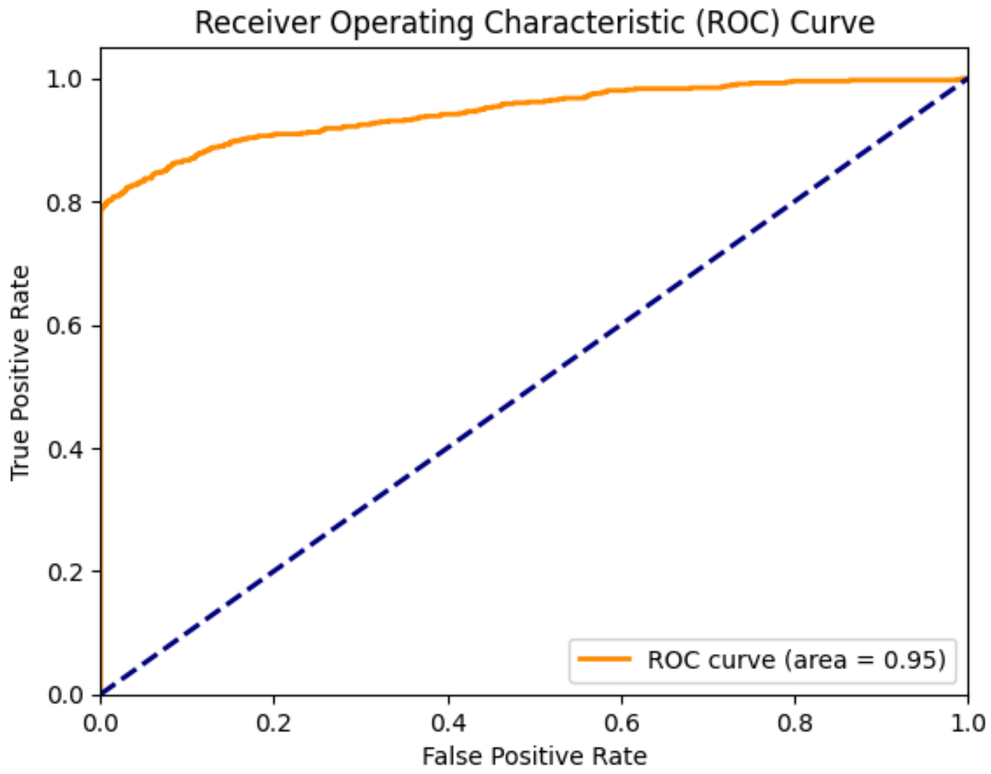
For the purpose of avoiding excessive usage of columns in resulting machine learning model, I ahev use Label encoder to reduce the number of unique values.

TRAINING MACHINE LEARNING MODEL



Random Forest Model

This machine learning model performed well in the given task and returned 0.78 auc metric. The model is trained with over 200000 data from the dataset.



Tabular model of fastai

This machine learning model performed far better than Random forest and demonstrated 0.95 auc metric. The model is trained using only 3% of the data and test with other 100000 data.

THANK YOU

FOR YOUR ATTENTION

Umidjon Sattorov(student at Skillbox platform)

sattorov7474@gmail.com