



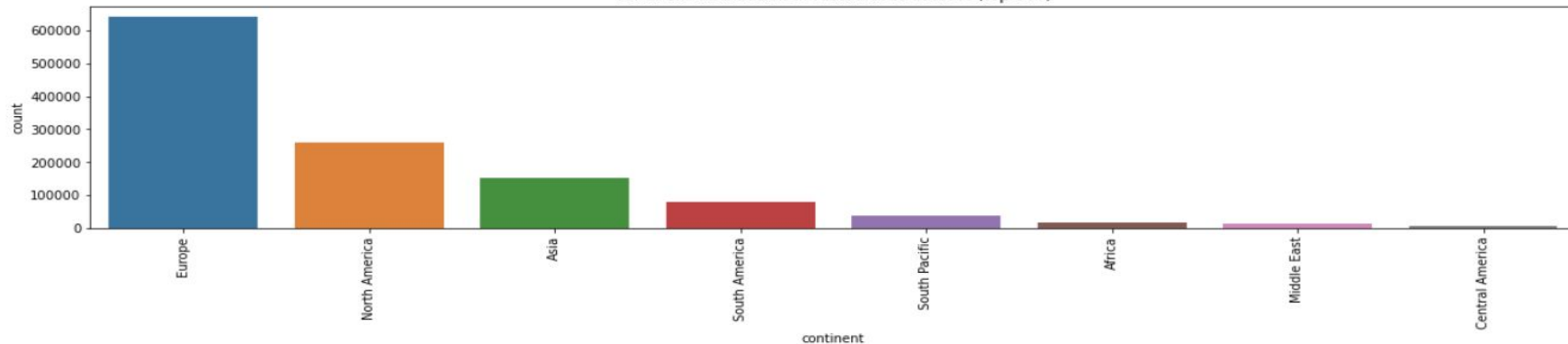
Premium clients prediction



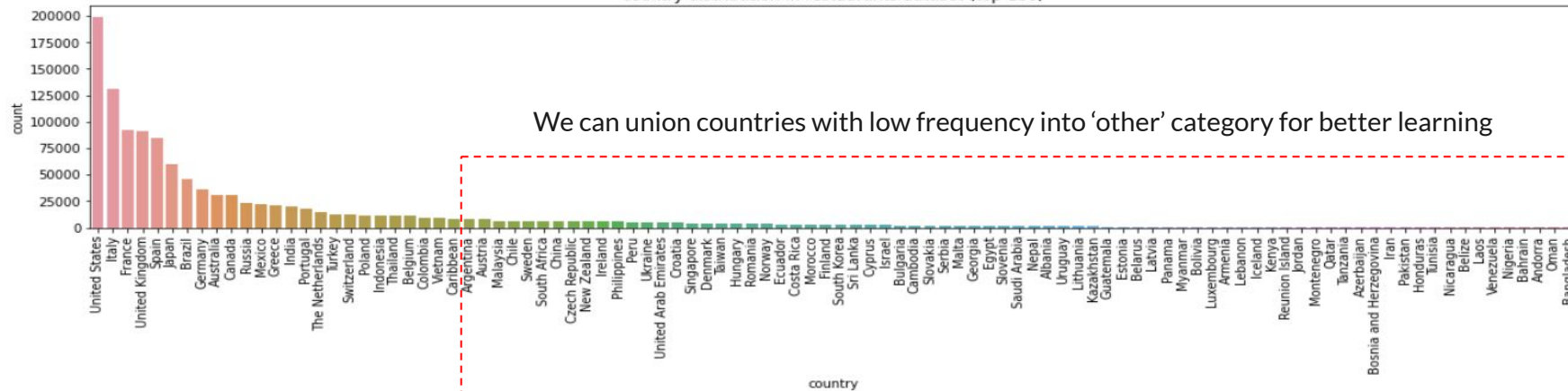
Exploratory data analysis. Restaurants dataset

EDA: Categorical features

continent distribution in restaurants dataset (top-100)



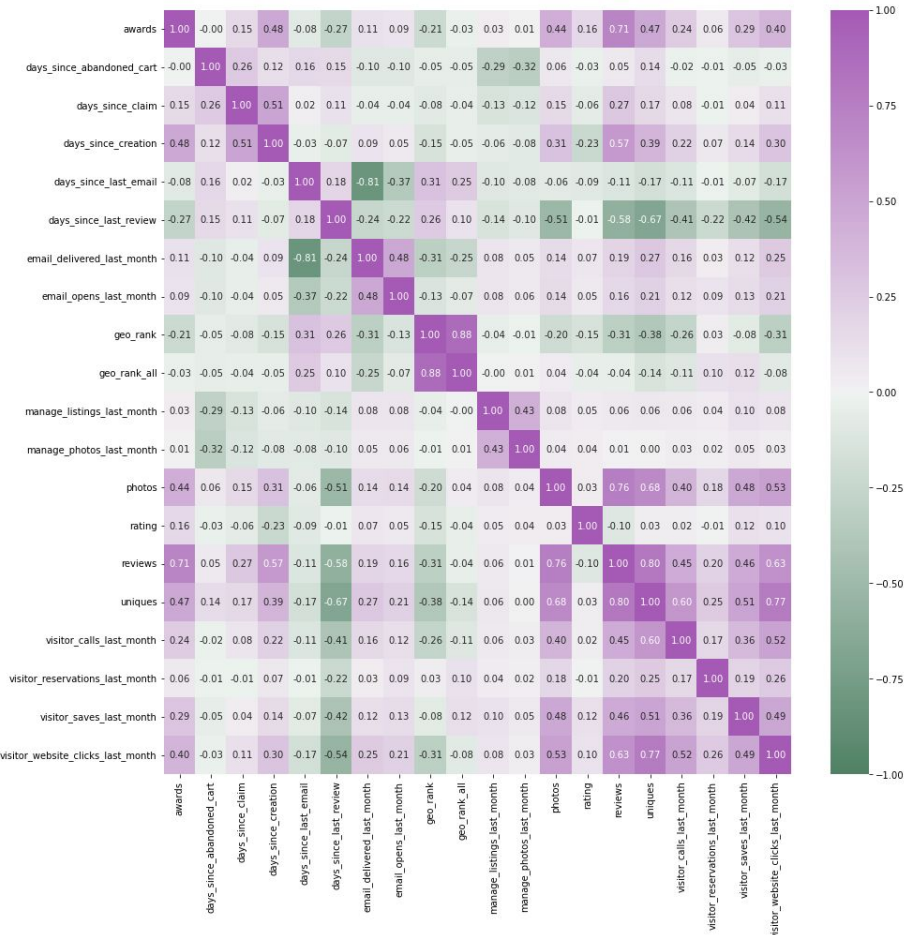
country distribution in restaurants dataset (top-100)



EDA: Numerical features. There a lot of NaNs and an outlier in days_since_claim which are processed in the solution

	count	mean	std	min	25%	50%	75%	max
awards	383926.00	3.67	2.53	1.00	1.00	3.00	5.00	11.00
days_since_abandoned_cart	10192.00	160.23	91.30	0.00	85.00	161.00	221.00	365.00
days_since_claim	1212469.00	1339.62	869.81	-1.00	698.00	1146.00	1905.00	4616.00
days_since_creation	1191427.00	2310.74	1268.04	1.00	1264.00	2248.00	3149.00	5800.00
days_since_last_email	863440.00	30.36	42.89	1.00	11.00	17.00	21.00	276.00
days_since_last_review	1111006.00	394.61	478.77	1.00	49.00	152.00	678.00	5474.00
email_delivered_last_month	863420.00	3.76	3.93	0.00	1.00	3.00	5.00	606.00
email_opens_last_month	863420.00	0.93	1.81	0.00	0.00	0.00	1.00	66.00
geo_rank	1115935.00	619.98	1764.71	0.00	10.00	53.00	324.00	29151.00
geo_rank_all	1212469.00	1387.97	3092.44	0.00	32.00	162.00	974.00	18085.00
manage_listings_last_month	1091952.00	0.23	1.90	0.00	0.00	0.00	0.00	207.00
manage_photos_last_month	1091952.00	0.08	1.09	0.00	0.00	0.00	0.00	201.00
photos	1102309.00	61.74	152.00	0.00	9.00	23.00	59.00	23964.00
rating	1009577.00	8.44	1.24	2.00	8.00	9.00	9.00	10.00
reviews	1111009.00	133.85	328.61	0.00	10.00	38.00	129.00	38758.00
uniques	1212117.00	1237.86	2586.53	1.00	128.00	443.00	1281.00	371437.00
visitor_calls_last_month	1205651.00	1.53	5.38	0.00	0.00	0.00	1.00	666.00
visitor_reservations_last_month	1205651.00	0.66	5.37	0.00	0.00	0.00	0.00	921.00
visitor_saves_last_month	1205651.00	0.89	4.83	0.00	0.00	0.00	0.00	741.00
visitor_website_clicks_last_month	1205651.00	8.62	26.44	0.00	0.00	1.00	7.00	1730.00

EDA: Numerical features. Correlation plot



1. Top-3 positive Spearman correlation:
 - uniques and reviews
 - visitor_website_clicks_last_month and uniques
 - photos and reviews
2. Top-3 negative Spearman correlation:
 - email_delivered_last_month and days_since_last_email
 - uniques and days_since_last_review
 - days_since_last_review and reviews



Exploratory data analysis. Subscription dataset

EDA: Subscription dataset

Customers choose to pay monthly more often than to pay annually



Subscription contains information about active users only

	location_id	is_active
count	7203.00	7203.00
mean	11870535.77	1.00
std	7702506.01	0.00
min	321605.00	1.00
25%	4080324.00	1.00
50%	12338380.00	1.00
75%	18955887.50	1.00
max	23869957.00	1.00

There are fewer premium customers than non-premium customers

		number	%
0	all restaurants	1212469	100.00%
1	restaurants with premium	6724	0.55%
2	premium, but not in the restaurants.csv	479	0.04%



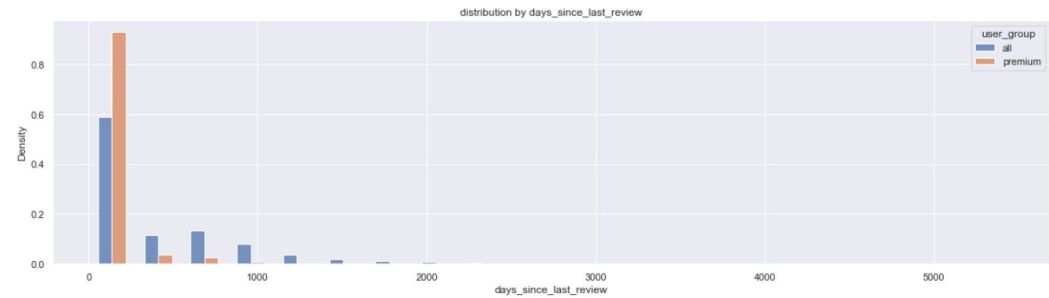
Exploratory data analysis. Explore Premium clients

Methodology:

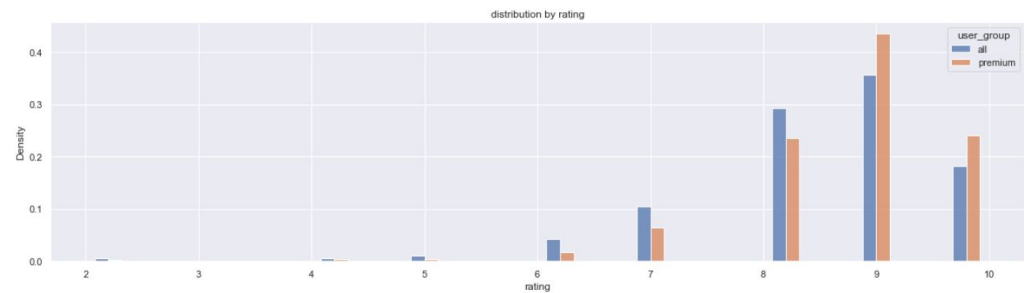
In the following slides, I **compare premium restaurants with the entire restaurants base**. I did this to understand the unique behaviors of premium customers and **to identify any distinct patterns**. Therefore, I created distributions for both premium restaurants and all restaurants based on different features.

EDA: Premium clients comparison (1/2)

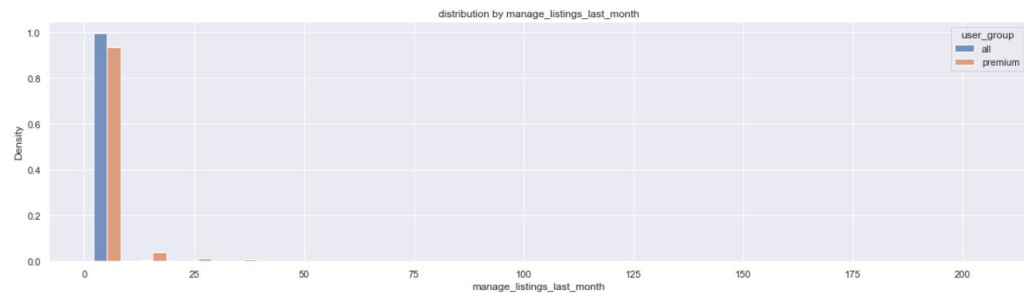
Premium clients have more fresh reviews



Premium clients have higher rating

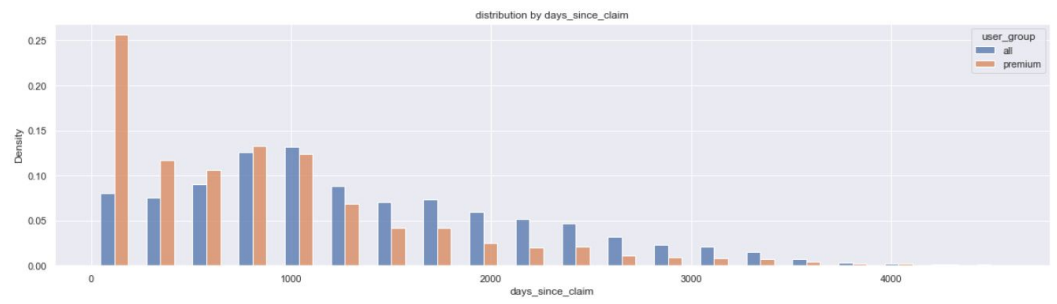


Premium clients update listing more often

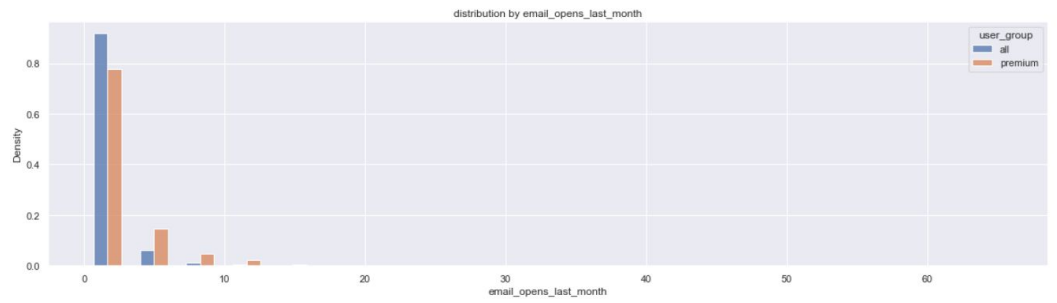


EDA: Premium clients comparison (2/2)

Premium clients claimed the listing on website more recently



Premium clients open emails more often



Premium clients have more redirected website clicks



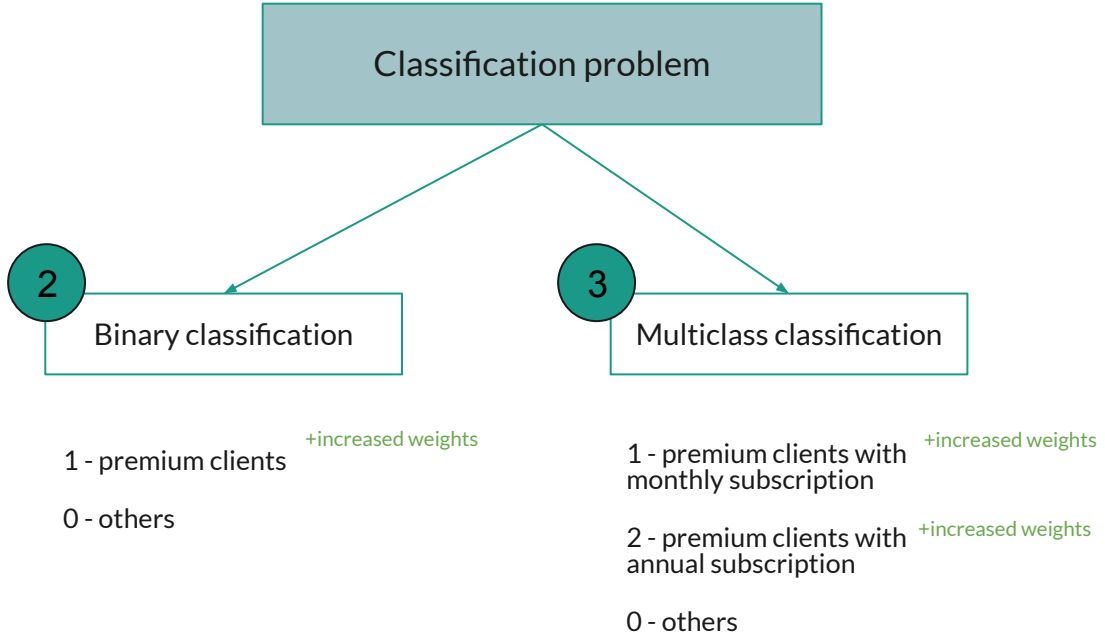


Model building process

Modelling: There are 3 possible solutions to this problem

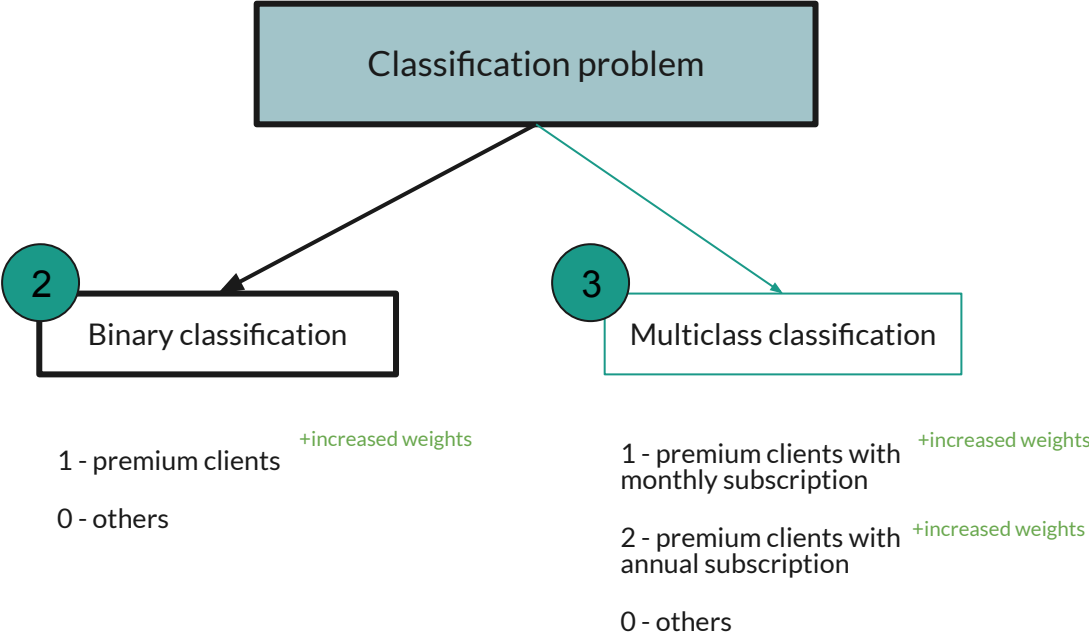
1

Find restaurants most similar to premium



Modelling: There are 3 possible solutions to this problem

1 Find restaurants most similar to premium



In a real situation, I would choose a few strategies and compare them with each other. Now I want to choose one strategy that is not too simple and not too complex, so I'm going to develop a solution for the binary classification task with Random Forest

What about weights?


I would like to increase the weights for class 1 or create an unbalanced dataset where class 1 is the majority class.

Why do I want to increase weights?


- 1) **I'm more certain about who the premium client is than about those who are not.** Among the 0 class, there are restaurants that could be premium if they knew about this option. However, among premium clients, there are no non-premium clients.
- 2) **If there are two similar clients in the train set and one has premium status** while the other does not, I want my model to be trained to consider **both of these clients as premium.**


Modelling: feature engineering

	count	mean	std	min	25%	50%	75%	max
awards	383926.00	3.67	2.53	1.00	1.00	3.00	5.00	11.00
days_since_abandoned_cart	10192.00	160.23	91.30	0.00	85.00	161.00	221.00	365.00
days_since_claim	1212469.00	1339.62	869.81	-1.00	698.00	1146.00	1905.00	4616.00
days_since_creation	1191427.00	2310.74	1268.04	1.00	1264.00	2248.00	3149.00	5800.00
days_since_last_email	863440.00	30.36	42.89	1.00	11.00	17.00	21.00	276.00
days_since_last_review	1111006.00	394.61	478.77	1.00	49.00	152.00	678.00	5474.00
email_delivered_last_month	863420.00	3.76	3.93	0.00	1.00	3.00	5.00	606.00
email_opens_last_month	863420.00	0.93	1.81	0.00	0.00	0.00	1.00	66.00
geo_rank	1115935.00	619.98	1764.71	0.00	10.00	53.00	324.00	29151.00
geo_rank_all	1212469.00	1387.97	3092.44	0.00	32.00	162.00	974.00	18085.00
manage_listings_last_month	1091952.00	0.23	1.90	0.00	0.00	0.00	0.00	207.00
manage_photos_last_month	1091952.00	0.08	1.09	0.00	0.00	0.00	0.00	201.00
photos	1102309.00	61.74	152.00	0.00	9.00	23.00	59.00	23964.00
rating	1009577.00	8.44	1.24	2.00	8.00	9.00	9.00	10.00
reviews	1111009.00	133.85	328.61	0.00	10.00	38.00	129.00	38758.00
uniques	1212117.00	1237.86	2586.53	1.00	128.00	443.00	1281.00	371437.00
visitor_calls_last_month	1205651.00	1.53	5.38	0.00	0.00	0.00	1.00	666.00
visitor_reservations_last_month	1205651.00	0.66	5.37	0.00	0.00	0.00	0.00	921.00
visitor_saves_last_month	1205651.00	0.89	4.83	0.00	0.00	0.00	0.00	741.00
visitor_website_clicks_last_month	1205651.00	8.62	26.44	0.00	0.00	1.00	7.00	1730.00

 - fill NaN with -1000 (since tree-based approach and I can NaN replace with extreme value)

 - fill NaN with 0

 - fill NaN with days_since_claim

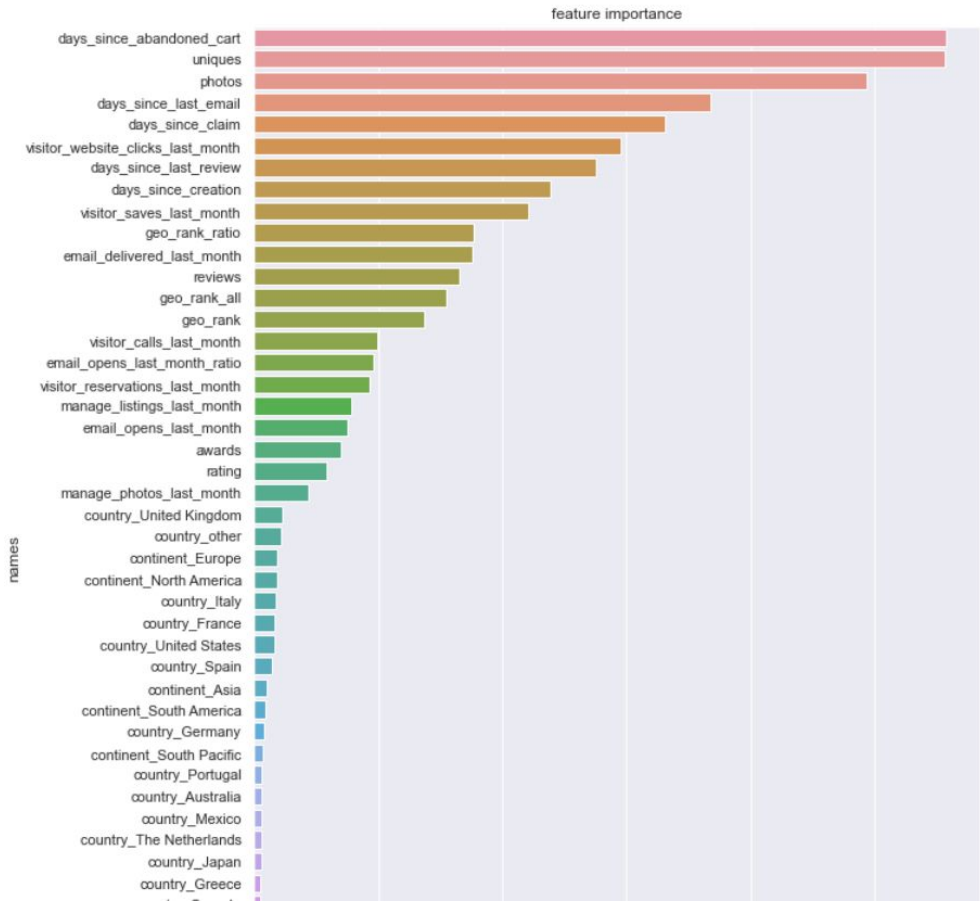
 - fill NaN in this way:
1) calculated ratio = geo_rank/geo_rank_all
2) calculate mean_ratio = average(ratio)
3) fill it with mean_ratio * geo_rank_all

Added 2 features:

- 1) **geo_rank_ratio** = **geo_rank/geo_rank_all** - it can show how good restaurant is
- 2) **email_opens_last_month_ratio** = **email_opens_last_month/email_delivered_last_month** - it shows if owner is interested in any marketing communications

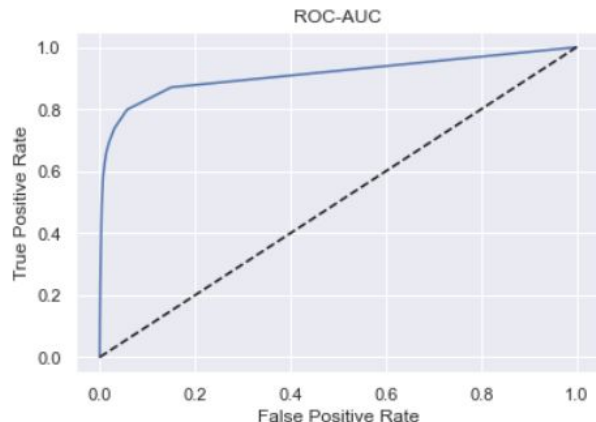
The final model: feature importance

I have chosen **RandomForest** and increased the weights for the class 1 by two times.



The final model: performance evaluation

ROC-AUC shows the quality of classification.



RoC-AUC of my model is 91% and its considered good result

Ranking metrics: NDCG@k, precision@k and recall@k

1 class in test: 1345

ndcg@50: 0.635125183826517
precision@50: 0.66
recall@50: 0.02453531598513011

1 class in test: 1345

ndcg@1345: 0.44373540389771293
precision@1345: 0.42230483271375463
recall@1345: 0.42230483271375463

NDCG@k and **precision@k** help to understand if the model correctly ranked premium restaurants at the top.

The **recall@k** metric is more interesting to me, and I would like it to be chosen as the main metric. This is because if the sales team calls the top-k restaurants, I would like to have as many premium clients as possible as the result.



Next steps

- 1) I would like to experiment with the multiplier for class weights. Once we agree on the main metric, I can optimize it by increasing or decreasing the weights accordingly.
- 2) Additionally, I would like to explore other models to see their performance.
- 3) Finally, for the best model, we can run an AB test to evaluate its effectiveness.