X Education Lead Score Prediction

X Education

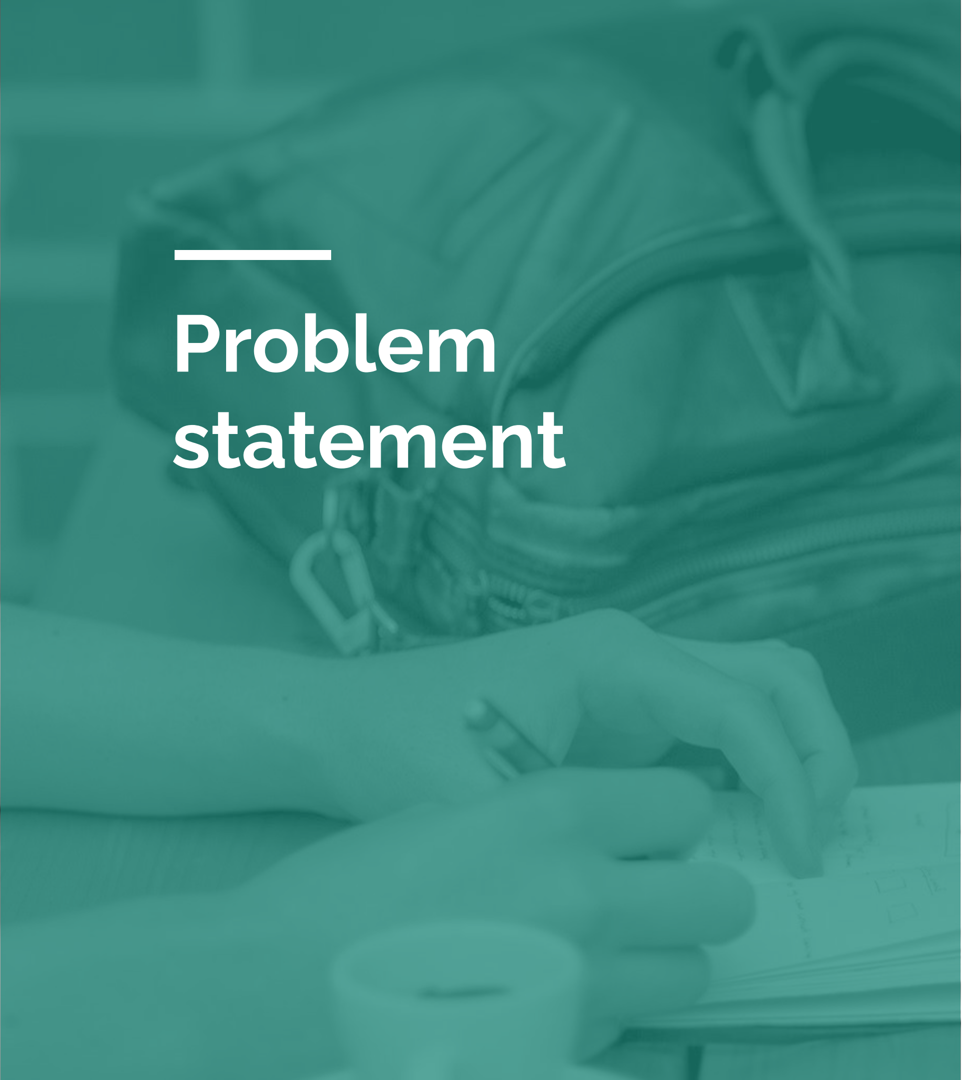
- **Product:** Website with online courses for industry professionals

- **Marketing channels:** several websites and search engines like Google

- **Leads:** People who fill up a form providing their email address or phone number on the X Education website

- **Sales process:** Sales Team calls or emails Leads to convert them to paying customers.

- **Conversion Rate:** around 30%
  **according to X Education*
  **dataset= 39%*



Problem statement

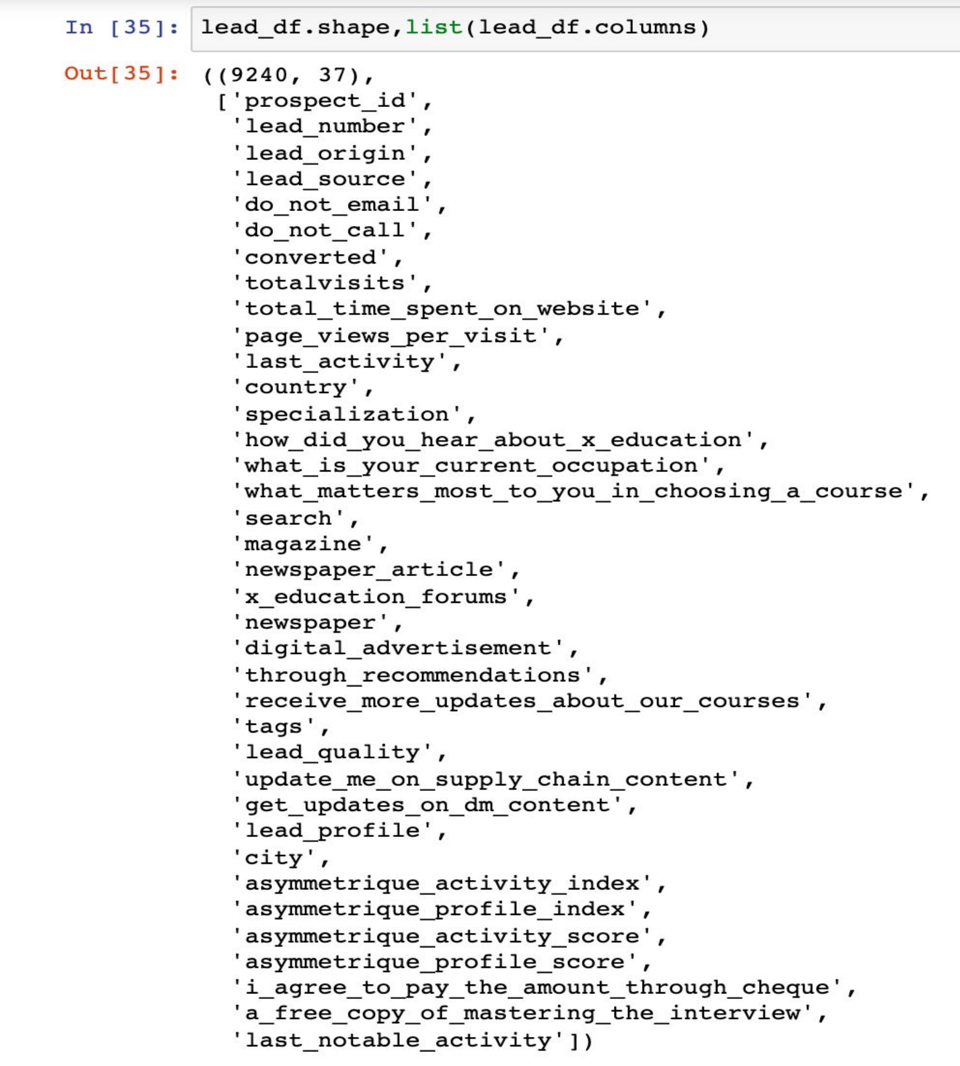
- X Education receives many leads, but the conversion rate is very low: around 30%.

- Sales team lacks an effective lead prioritisation system.

Dataset

Lead list from the sales team including lead conversion

**source: kaggle*



```
In [35]: lead_df.shape,list(lead_df.columns)

Out[35]: ((9240, 37),
          ['prospect_id',
           'lead_number',
           'lead_origin',
           'lead_source',
           'do_not_email',
           'do_not_call',
           'converted',
           'totalvisits',
           'total_time_spent_on_website',
           'page_views_per_visit',
           'last_activity',
           'country',
           'specialization',
           'how_did_you_hear_about_x_education',
           'what_is_your_current_occupation',
           'what_matters_most_to_you_in_choosing_a_course',
           'search',
           'magazine',
           'newspaper_article',
           'x_education_forums',
           'newspaper',
           'digital_advertisement',
           'through_recommendations',
           'receive_more_updates_about_our_courses',
           'tags',
           'lead_quality',
           'update_me_on_supply_chain_content',
           'get_updates_on_dm_content',
           'lead_profile',
           'city',
           'asymmetrique_activity_index',
           'asymmetrique_profile_index',
           'asymmetrique_activity_score',
           'asymmetrique_profile_score',
           'i_agree_to_pay_the_amount_through_cheque',
           'a_free_copy_of_mastering_the_interview',
           'last_notable_activity'])
```

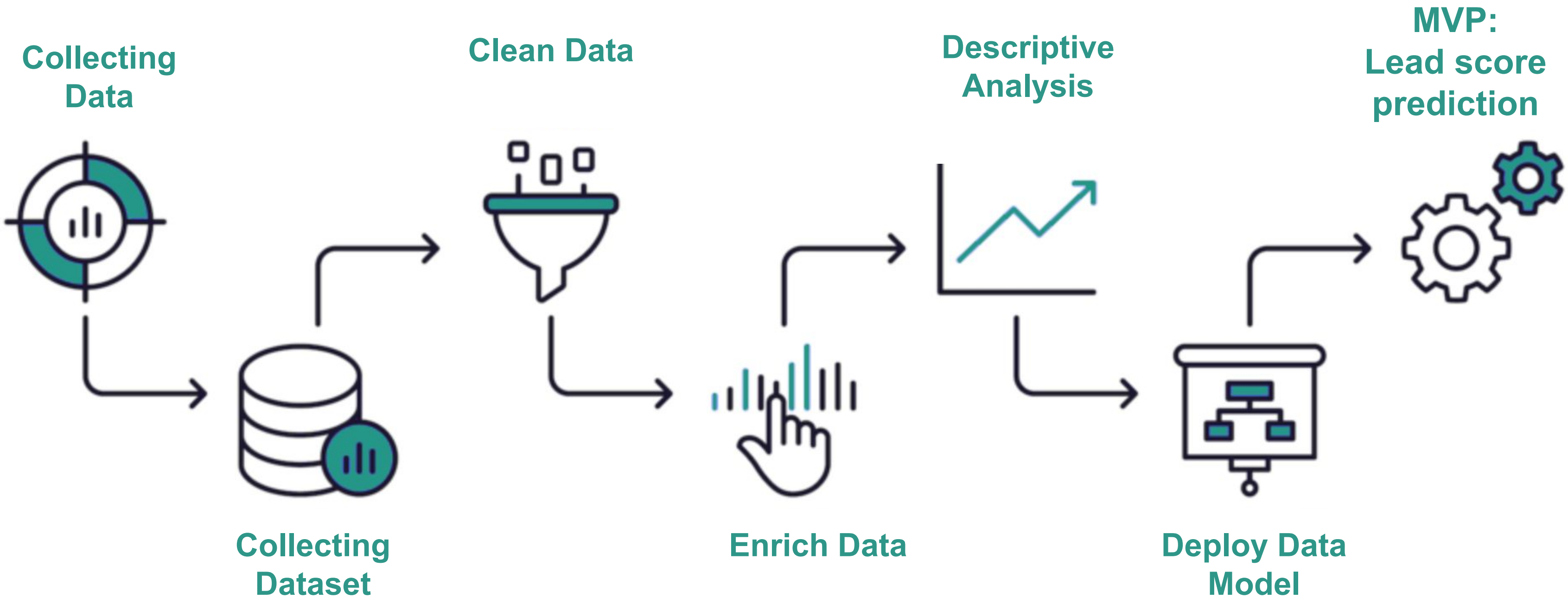
Objective

1. Conversion improvement proposals based on descriptive analysis

2. Lead scoring model which predicts the conversion probability

   (CEO wants focus on leads with 80% conversion probability)

Process



Collecting Data

Collecting Dataset

Clean Data

Enrich Data

Descriptive Analysis

Deploy Data Model

MVP: Lead score prediction

Data cleaning & processing

Original Dataset:

Columns : 37

Rows: 9239

Columns with missing data: 17

Methods

- Dropping columns with no variance
- &missing values>30%
- Data bucketing
- Outlier handling with Z-score
- Scaling numerical features
- Encoding categorical features
- Checking high correlation

Cleaned & processed Dataset:

Independent variables: 38 encoded and scaled columns

Dependent variable: binary

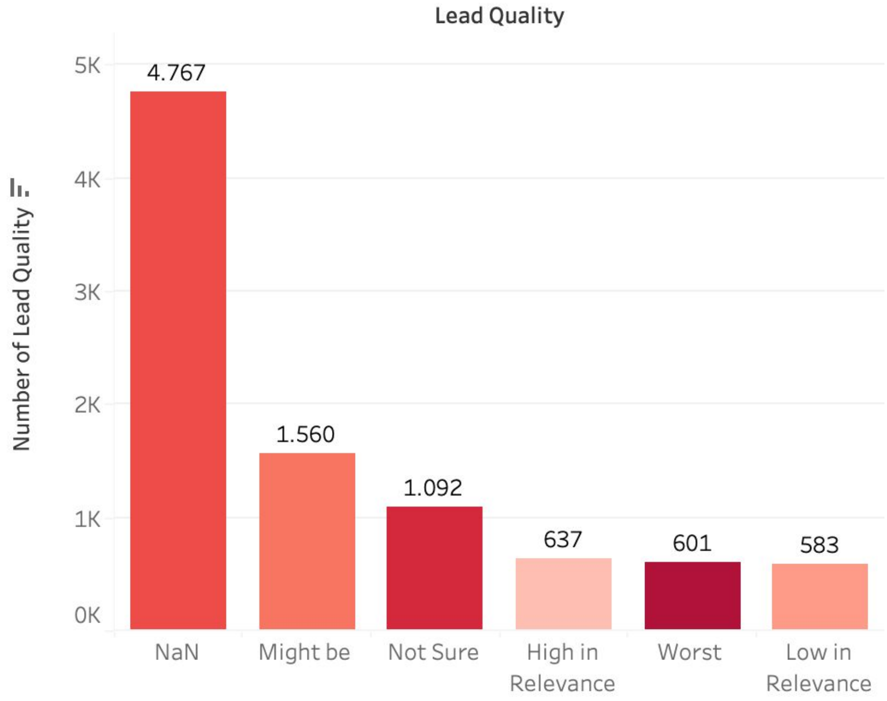
Rows: 9037 (98 %)

Columns with missing data: 0

1. Conversion improvement proposals based on descriptive analysis

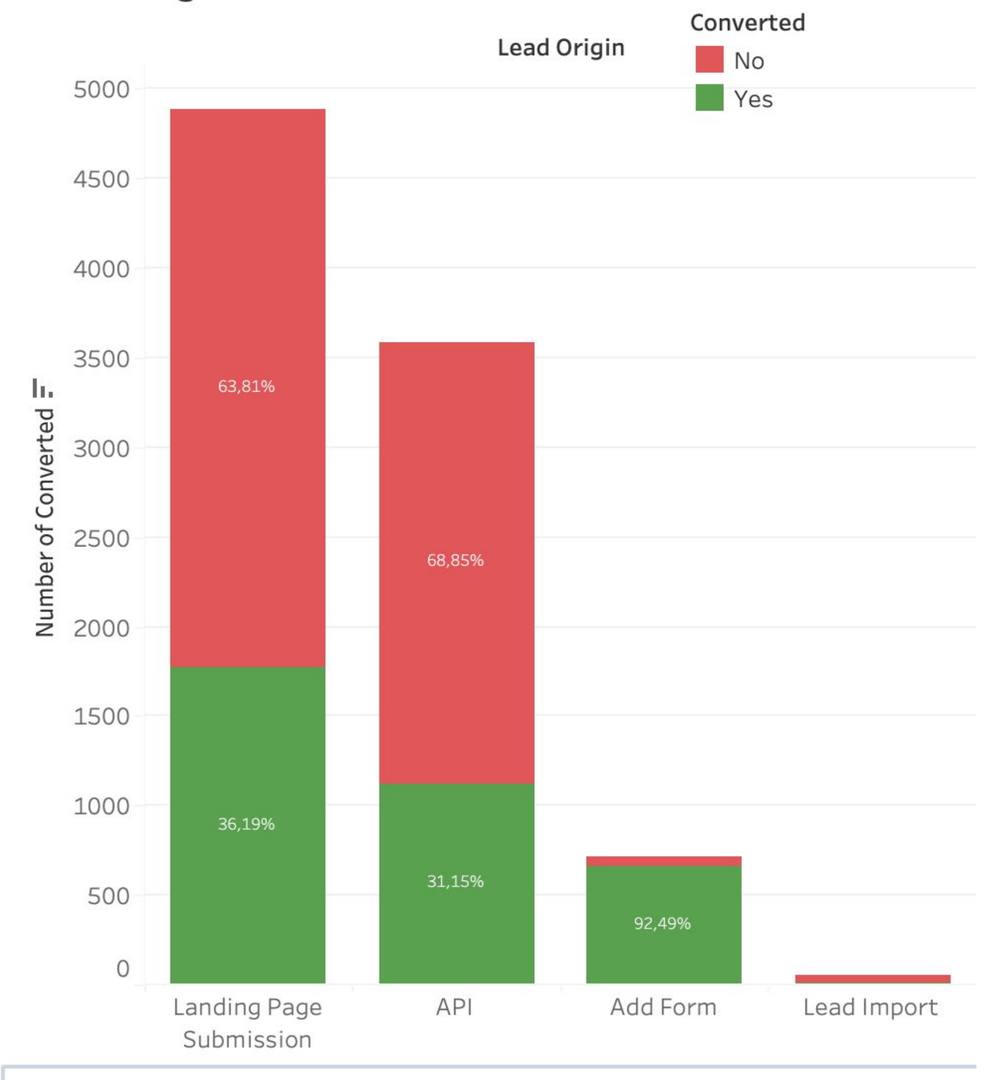
Status quo lead analysis

→ No established standard of understanding which leads are most likely to convert

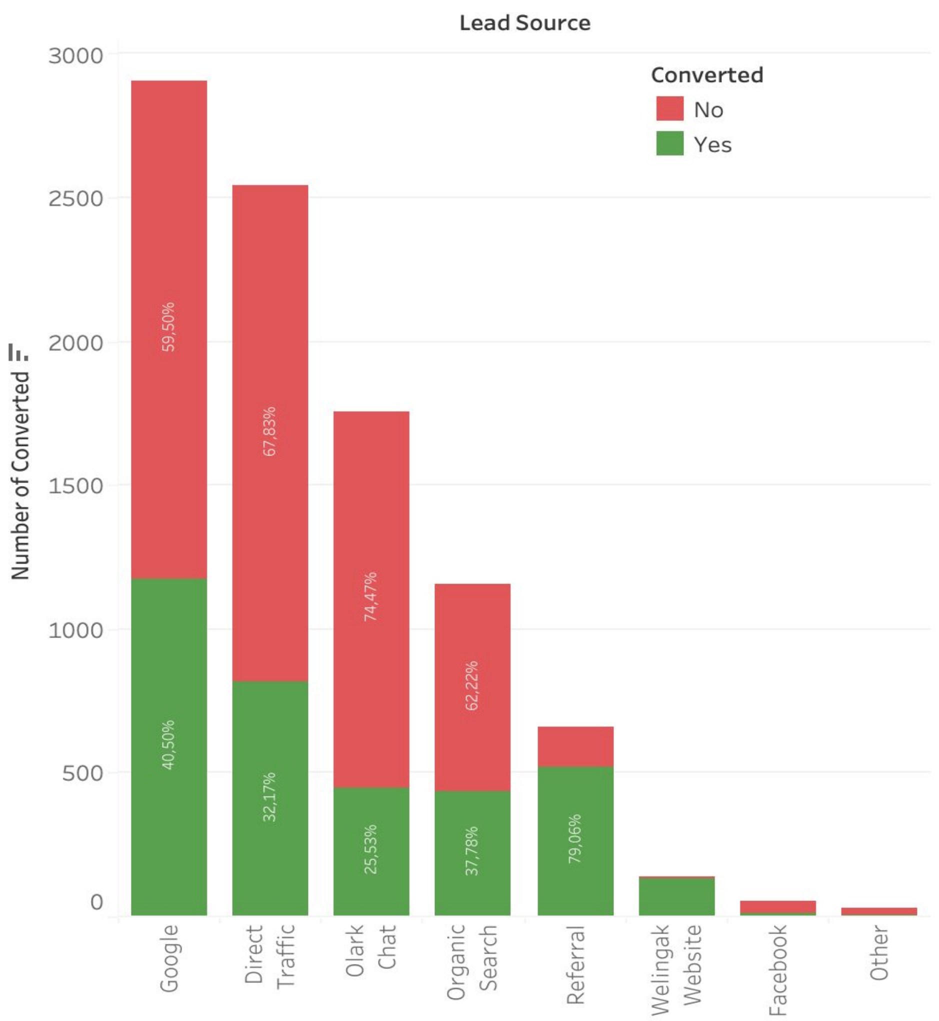


Indicates the quality of lead based on the data and intuition the employee who has been assigned to the lead.

Lead Origin

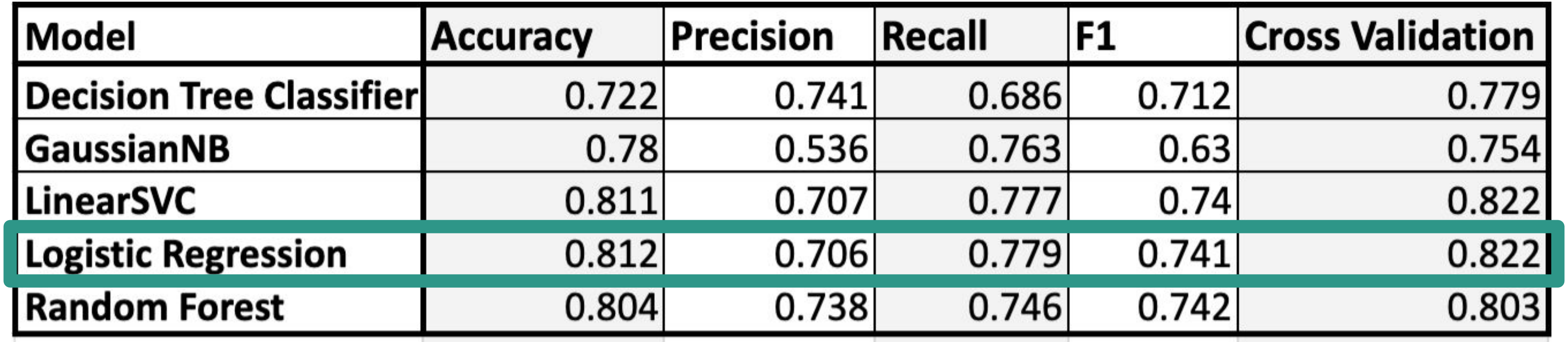

Converted
🟥 No
🟩 Yes
Lead Origin
Number of Converted
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
63,81%
36,19%
68,85%
31,15%
92,49%
Landing Page Submission
API
Add Form
Lead Import

Lead Source


Lead Source
Converted
No
Yes
Number of Converted
3000
2500
2000
1500
1000
500
0
59,50%
40,50%
67,83%
32,17%
74,47%
25,53%
62,22%
37,78%
79,06%
Google
Direct Traffic
Olark Chat
Organic Search
Referral
Welingak Website
Facebook
Other

2. Lead scoring model

Model selection

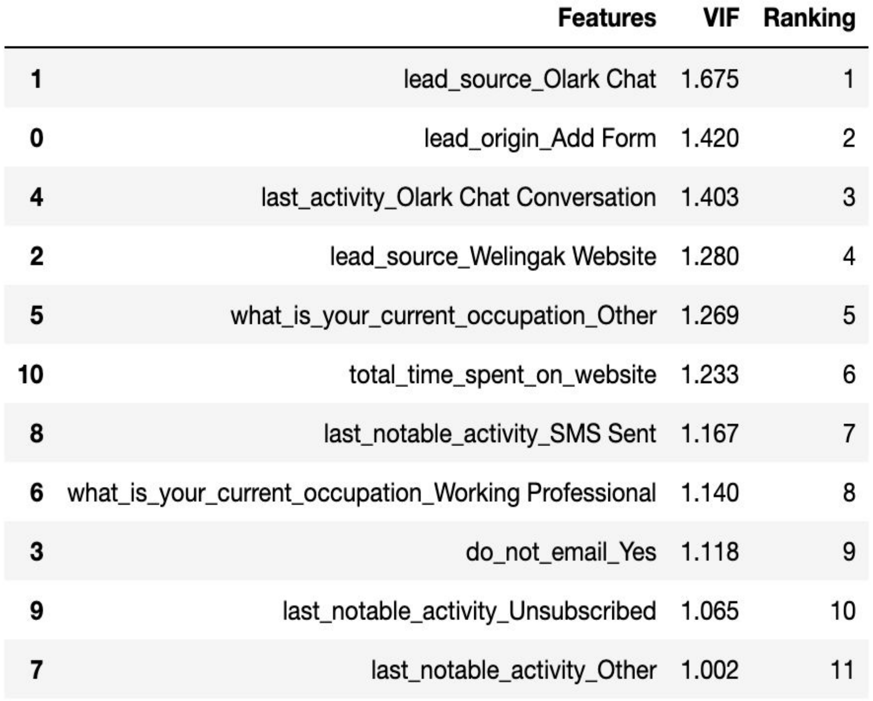


Model	Accuracy	Precision	Recall	F1	Cross Validation
Decision Tree Classifier	0.722	0.741	0.686	0.712	0.779
GaussianNB	0.78	0.536	0.763	0.63	0.754
LinearSVC	0.811	0.707	0.777	0.74	0.822
Logistic Regression	0.812	0.706	0.779	0.741	0.822
Random Forest	0.804	0.738	0.746	0.742	0.803

Final selected Features with RFE

Cross validation of selected features:
0.814

Cross validation with all features:
0.822



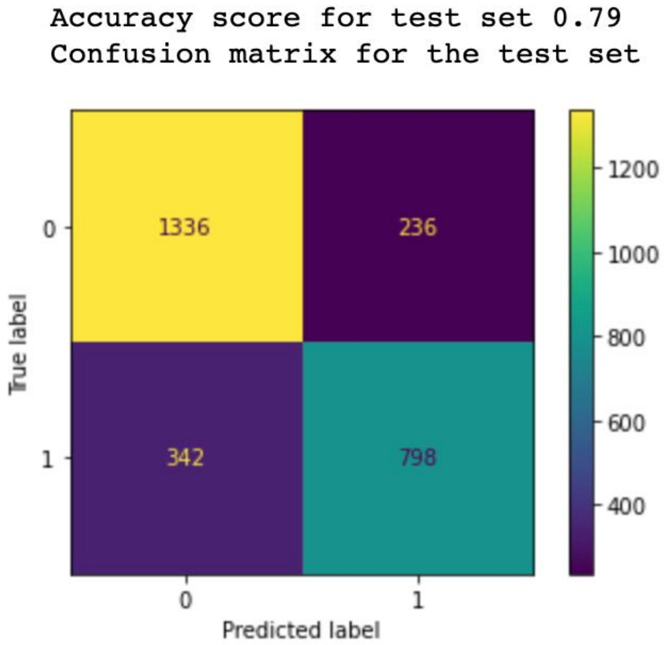
	Features	VIF	Ranking
1	lead_source_Olark Chat	1.675	1
0	lead_origin_Add Form	1.420	2
4	last_activity_Olark Chat Conversation	1.403	3
2	lead_source_Welingak Website	1.280	4
5	what_is_your_current_occupation_Other	1.269	5
10	total_time_spent_on_website	1.233	6
8	last_notable_activity_SMS Sent	1.167	7
6	what_is_your_current_occupation_Working Professional	1.140	8
3	do_not_email_Yes	1.118	9
9	last_notable_activity_Unsubscribed	1.065	10
7	last_notable_activity_Other	1.002	11

MVP
Lead Scoring prediction

-Allows Sales team to use a template to input lead details

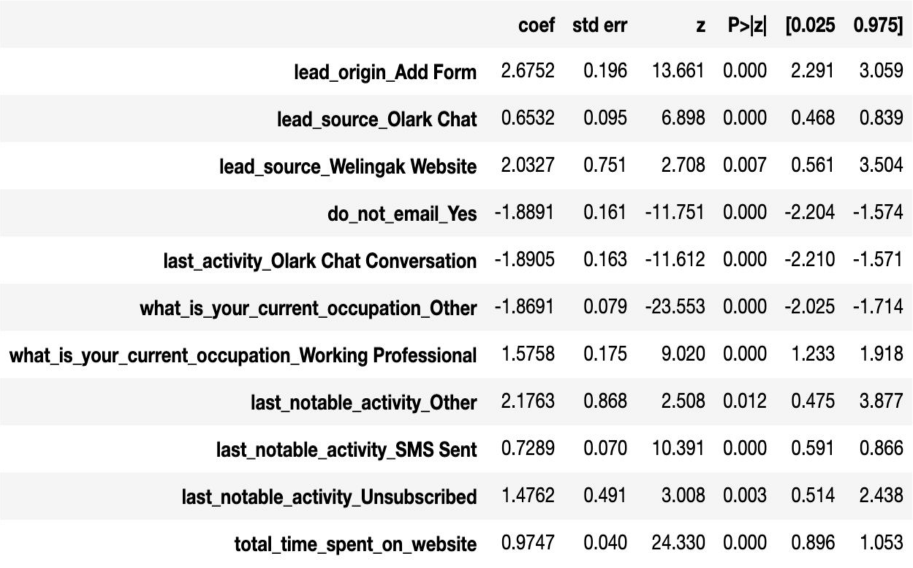
-Model outputs the conversion probability

Accuracy & Confusion matrix



```
Accuracy score for test set 0.79
Confusion matrix for the test set
```

Final selected Features analysis with statsmodels



	coef	std err	z	P>|z|	[0.025	0.975]
lead_origin_Add Form	2.6752	0.196	13.661	0.000	2.291	3.059
lead_source_Olark Chat	0.6532	0.095	6.898	0.000	0.468	0.839
lead_source_Welingak Website	2.0327	0.751	2.708	0.007	0.561	3.504
do_not_email_Yes	-1.8891	0.161	-11.751	0.000	-2.204	-1.574
last_activity_Olark Chat Conversation	-1.8905	0.163	-11.612	0.000	-2.210	-1.571
what_is_your_current_occupation_Other	-1.8691	0.079	-23.553	0.000	-2.025	-1.714
what_is_your_current_occupation_Working Professional	1.5758	0.175	9.020	0.000	1.233	1.918
last_notable_activity_Other	2.1763	0.868	2.508	0.012	0.475	3.877
last_notable_activity_SMS Sent	0.7289	0.070	10.391	0.000	0.591	0.866
last_notable_activity_Unsubscribed	1.4762	0.491	3.008	0.003	0.514	2.438
total_time_spent_on_website	0.9747	0.040	24.330	0.000	0.896	1.053

Contact details

E-Mail: JuleTwelkemeier@gmail.com

LinkedIn: https://www.linkedin.com/in/juletwelkemeier/

Github: https://github.com/JuleTwelkemeier/Final_project_ironhack