CASE STUDY
Data Scientist - Ranking

Simón Ramírez Hinestrosa

***For a detailed explanation of the process please refer to the Jupyter Notebooks**

Summary of the process

- Data cleaning and formatting
- Definition of the variables
- Model exploration
- Improvements

Data cleaning and formatting

- Explored the data, finding missing information
- Imputed values to missing rows in “session_duration”
- Worked on a reduced sample of the dataset
- Assumed that the “path_id_set” can be replaced by the number of paths “n_ids”

```
For column n_ids
2     207954
1      46824
3      29236
4       6859
5       2272
6        972
7        418
8        238
9        112
10        73
11        57
12        33
13        22
16        12
14        11
15         9
18         4
20         4
17         3
22         2
23         2
39         1
25         1
31         1
42         1
Name: n_ids, dtype: int64
There are 25 unique values
```

Definition of the variables

Categorical variables

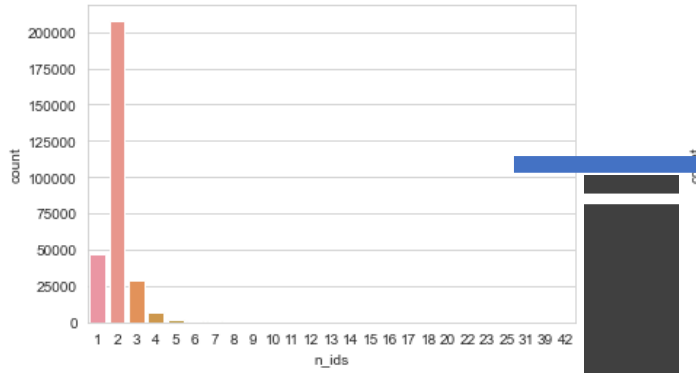
- Locale
- Traffic
- Agent Id
- Entry page
- n_ids

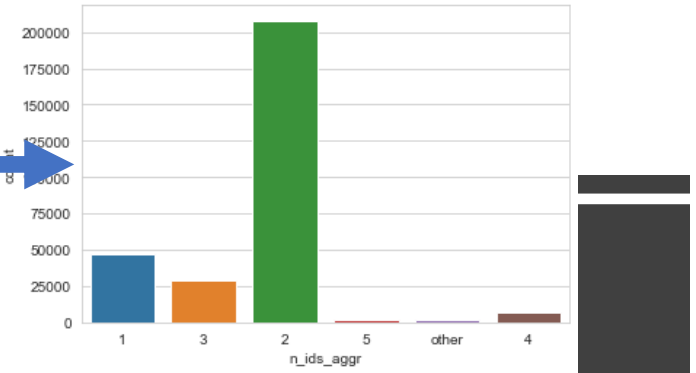
Aggregated variables with low counts

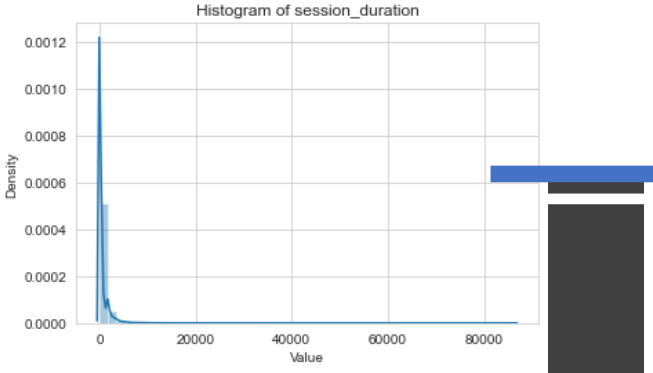
Numeric variables

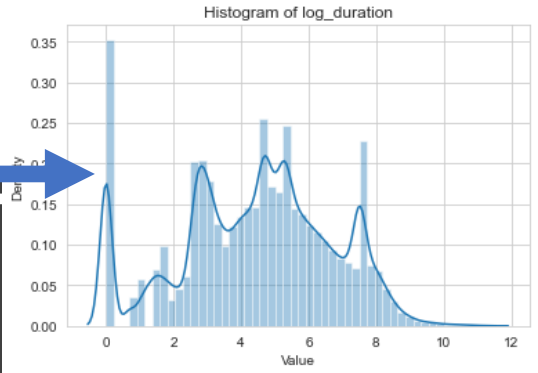
- Session duration
- Hour of day
- Day of week

Performed transformations to avoid skewed distributions



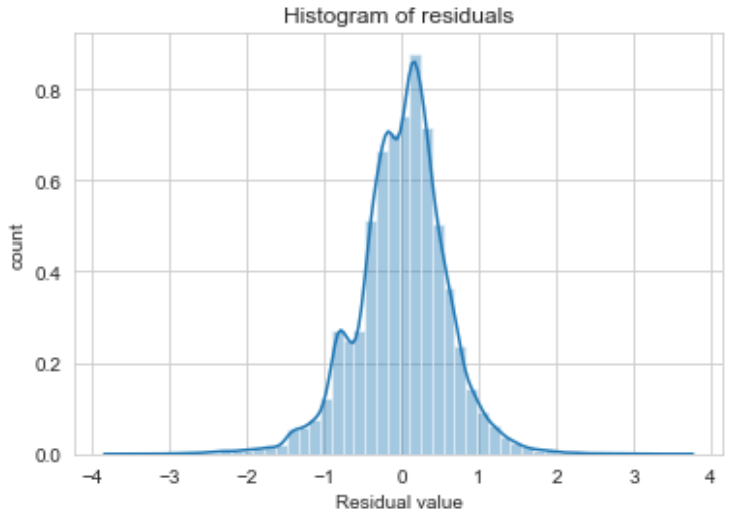






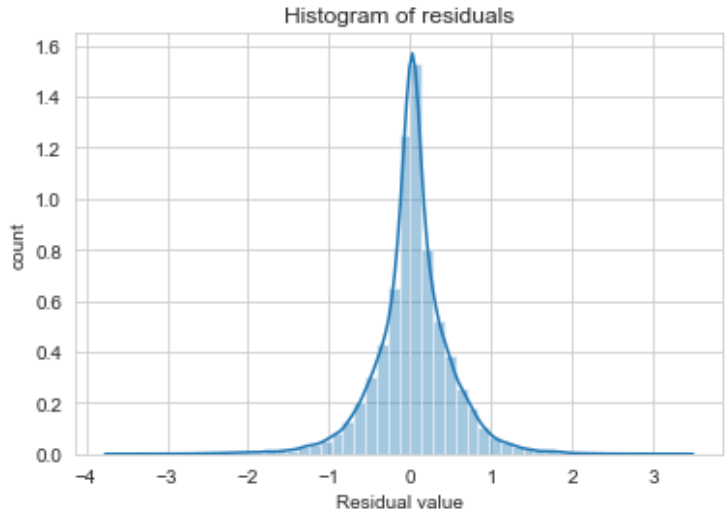
Model exploration

Linear regression




```
Mean Square Error      = 0.3341122898058772
Root Mean Square Error = 0.5780244716323671
Mean Absolute Error    = 0.43701277013234724
Median Absolute Error  = 0.3335552884109593
R^2                    = 0.7842941066816771
```

Neural network regressor




```
Mean Square Error      = 0.22599413990390207
Root Mean Square Error = 0.47538840951784056
Mean Absolute Error    = 0.32388040973287685
Median Absolute Error  = 0.20799367959915527
R^2                    = 0.8540961547358809
```

*The errors are for the logarithm of the number of hits

Improvements

- Explore the details of path_id_set
- Optimise NN parameters
- Bagging
- Remove outliners