Monthly Sales Predictive Model for New Leads

Michel Arruda - Nuvemshop test for data scientist position

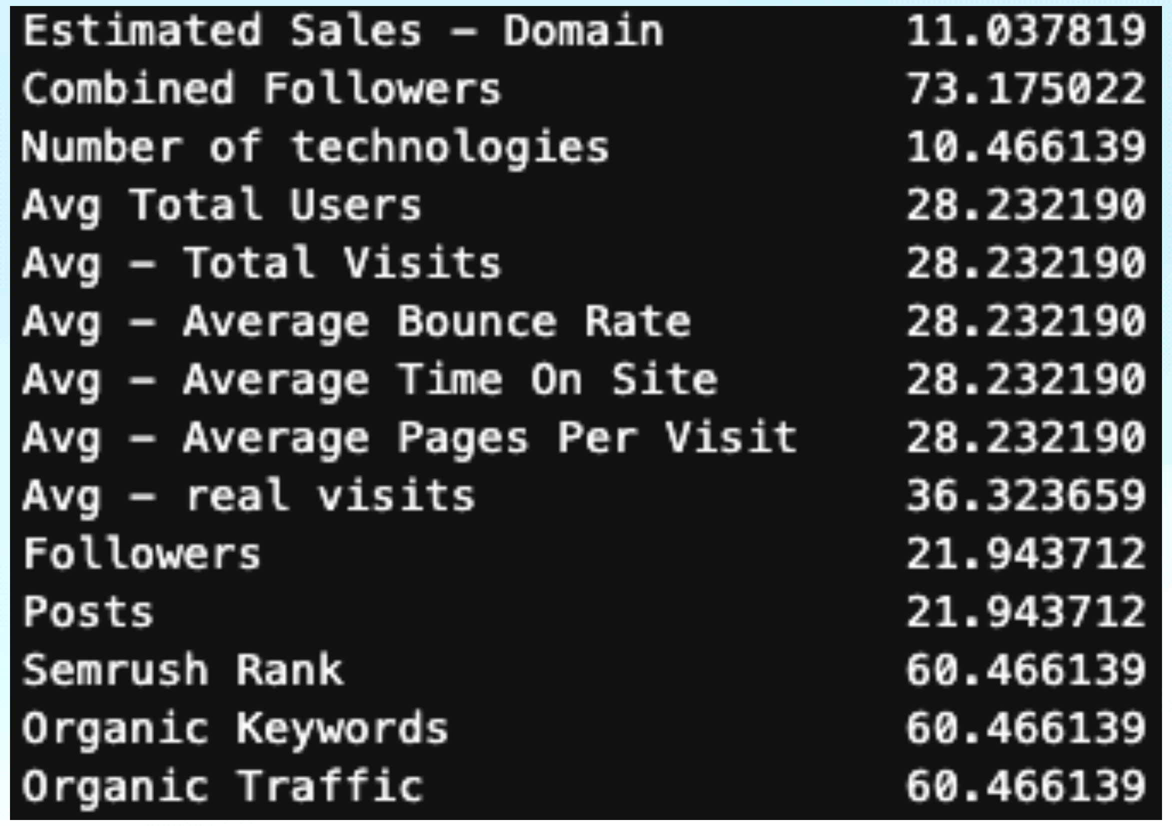
Introduction

- Objective: Predict monthly sales from new leads to optimize sales strategies.
- Problem: Determine which features are most relevant for predicting monthly sales and train a model for it.

Pre-processing and Analysis

Missing values

- Missing values in the features:



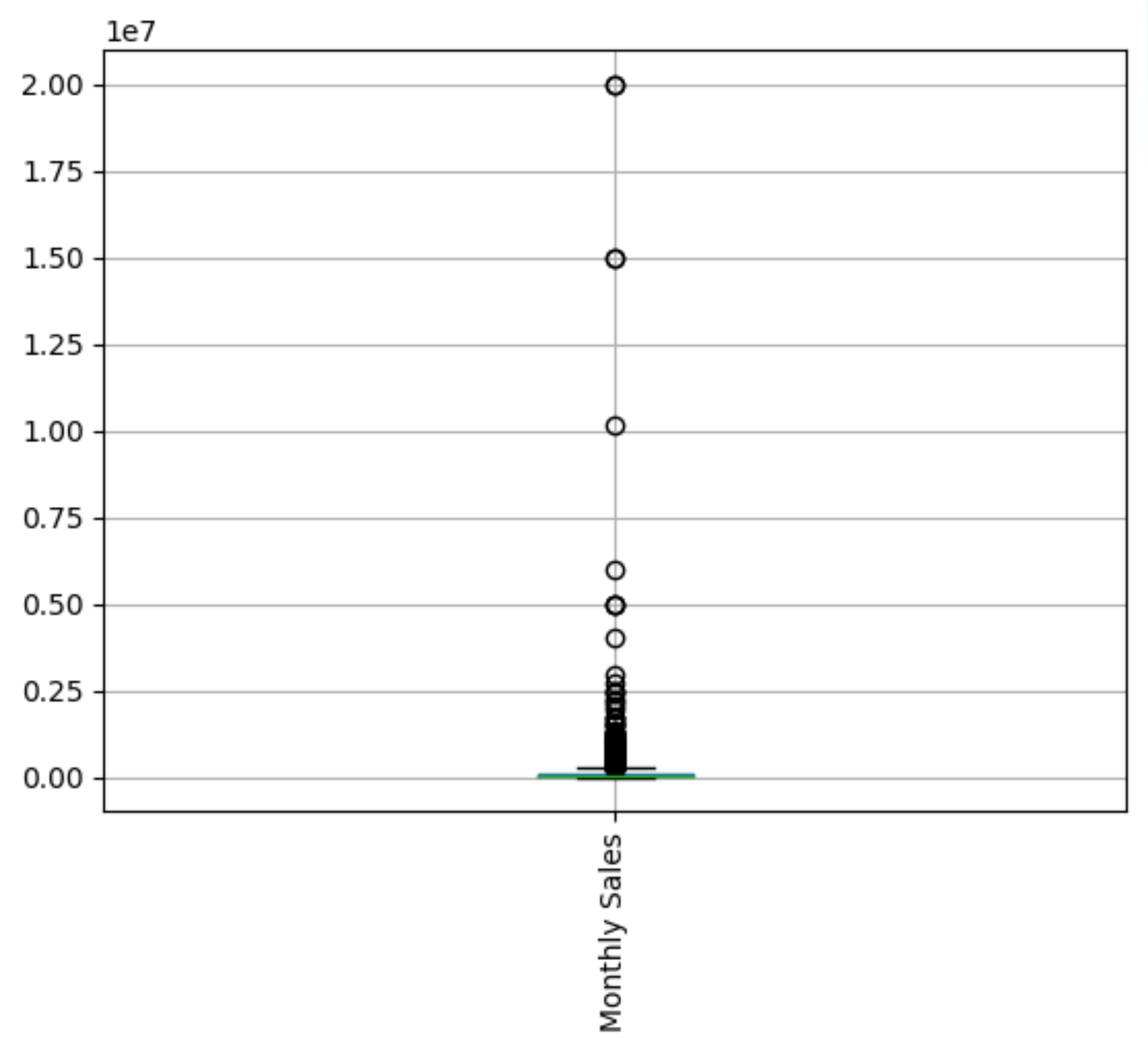
```
Estimated Sales - Domain        11.037819
Combined Followers              73.175022
Number of technologies          10.466139
Avg Total Users                 28.232190
Avg - Total Visits              28.232190
Avg - Average Bounce Rate       28.232190
Avg - Average Time On Site      28.232190
Avg - Average Pages Per Visit   28.232190
Avg - real visits               36.323659
Followers                       21.943712
Posts                           21.943712
Semrush Rank                    60.466139
Organic Keywords                60.466139
Organic Traffic                 60.466139
```

- Features removed with more than 50% missing values:
  - 'Employee Count’, 'Combined Followers’, 'Semrush Rank’, 'Organic Keywords','Organic Traffic','Adwords Keywords','Adwords Traffic','PLA keywords’
- The remaining features were treated using the median of the field values.

Pre-processing and Analysis

Outliers

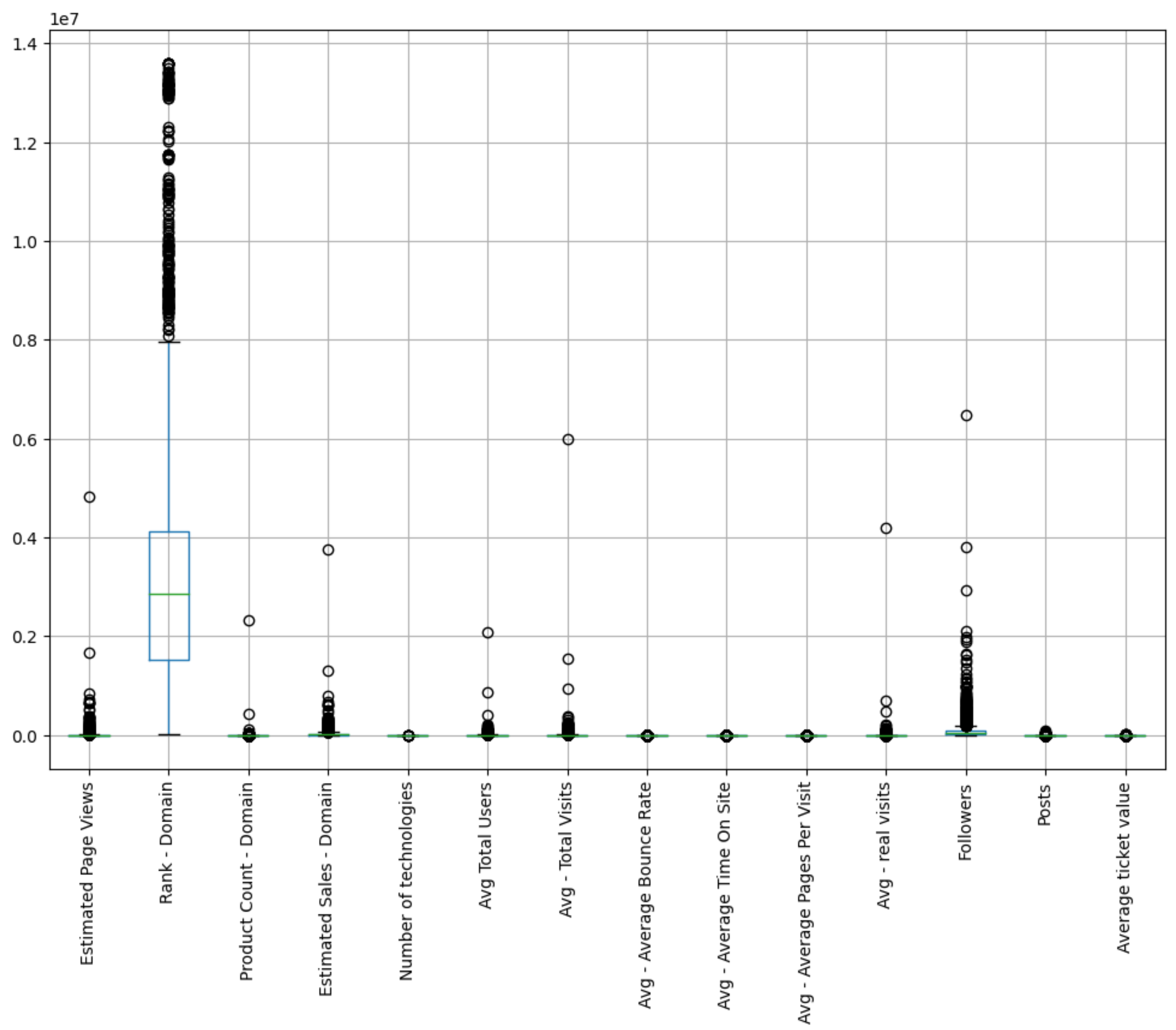
- The target has many outlier values. It can impact on the training model.



Pre-processing and Analysis

Outliers

- The features also have many outlier values.



Pre-processing and Analysis

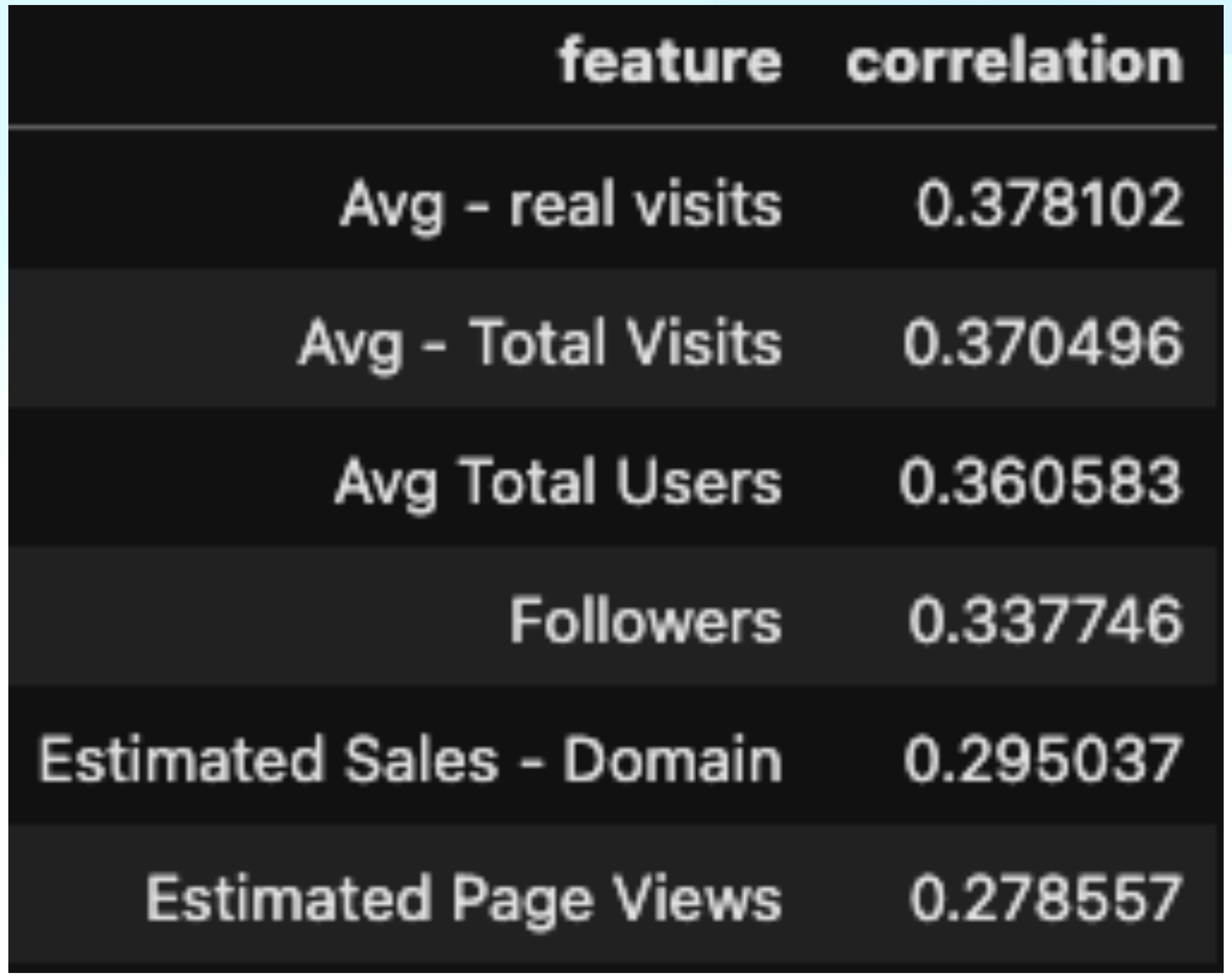
Outliers

- The outliers were removed for tests.
- When outliers were removed, the dataset had only 30% of the source rows.
- It was decided to keep the outliers for training.
- To lower the impact of outliers, the data was scaled

Pre-processing and Analysis

Correlation between features and target

- The most correlated features with the target are below:

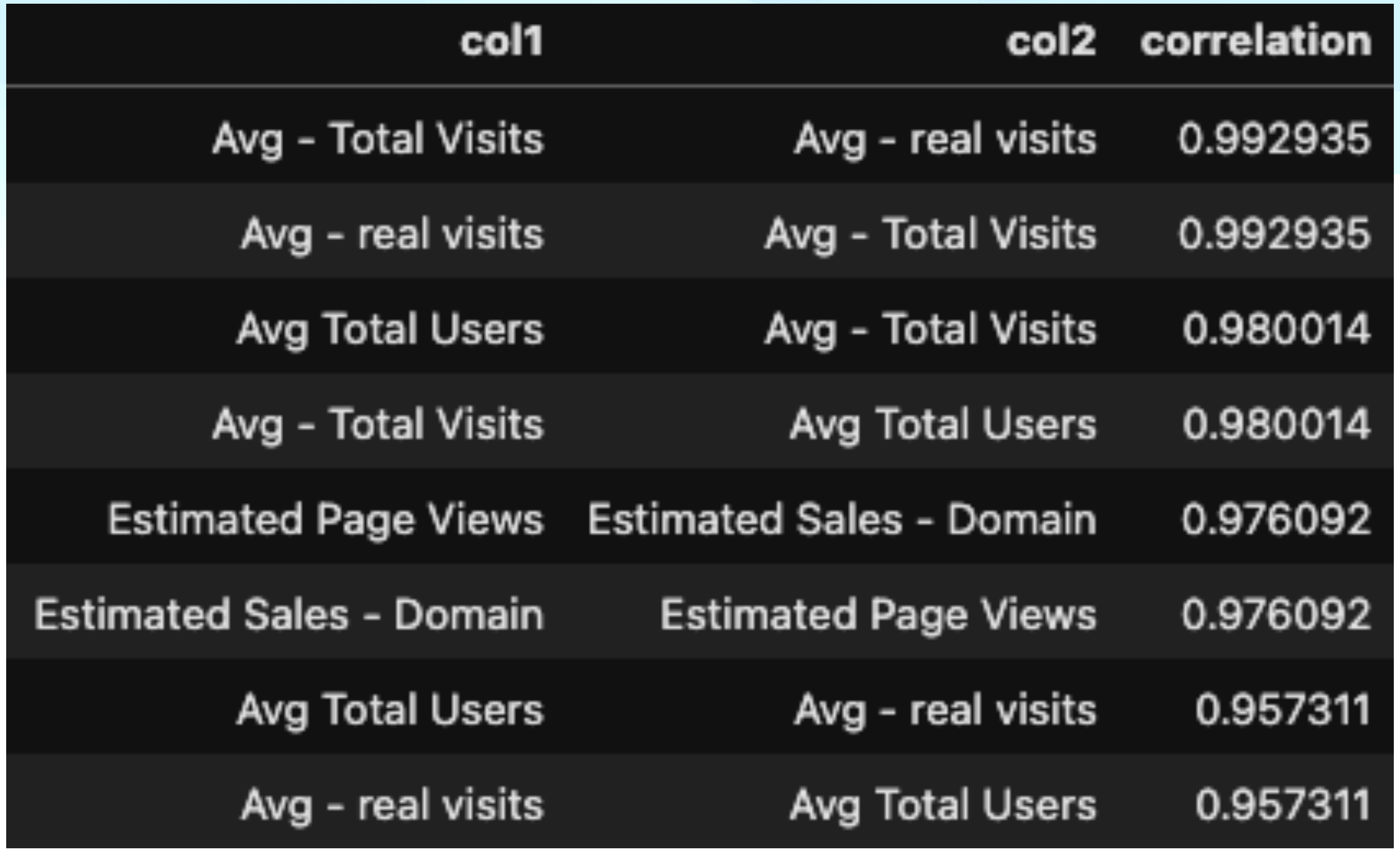


feature	correlation
Avg - real visits	0.378102
Avg - Total Visits	0.370496
Avg Total Users	0.360583
Followers	0.337746
Estimated Sales - Domain	0.295037
Estimated Page Views	0.278557

Pre-processing and Analysis

Correlation between features

- In the table, it can be seen that some of the features most correlated with the target are correlated with each other.
- To avoid collinearity problems, which could harm the model's training performance, they will be removed and only the most significant ones will be selected for training.
- Selected features for training:
  - 'Avg - real visits', 'Followers', 'Estimated Sales - Domain'



col1	col2	correlation
Avg - Total Visits	Avg - real visits	0.992935
Avg - real visits	Avg - Total Visits	0.992935
Avg Total Users	Avg - Total Visits	0.980014
Avg - Total Visits	Avg Total Users	0.980014
Estimated Page Views	Estimated Sales - Domain	0.976092
Estimated Sales - Domain	Estimated Page Views	0.976092
Avg Total Users	Avg - real visits	0.957311
Avg - real visits	Avg Total Users	0.957311

Model training

- List of applied model
  - Linear, Ridge, Lasso, ElasticNet, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, K-Nearest Neighbors, Support Vector Regressor, XGBRegressor
- The best model was the Decision Tree
- Metrics:
  - R^2 Score: 0.779151
  - RMSE: 3.601194e+05
- The model performed well in R^2, but the high RMSE value suggests that the target has distant values, impacting accuracy.

Lead Classification

- Additional objective: Classify predicted values as high, medium, or low to facilitate lead classification.
- A possible approach would be:
  - High: Sales above a certain threshold.
  - Medium: Sales in an intermediate range.
  - Low: Sales below a certain threshold.

Lead Classification

- For the threshold, it was used the percentiles (25% and 75%)
- Below Percentile 25% = Low
  - *value =< 24416.92 = 'low'*
- Between Percentile 25% and Percentile 75% = Medium
  - *value > 24416.92 and value < 121368.77 = 'medium'*
- Above of Percentile 75% = High
  - *value >= 121368.77 = 'high'*

Next Steps

- Remove outliers and train model to evaluate it
- Try to create new features and evaluate on training
- Try to improve the model in Hyperparameter Tuning

Conclusion

- The Decision Tree model showed good overall performance.
- Need for fine adjustments due to the impact of distant values on the target.
- Lead classification will be useful for strategic actions.