Analysis of Site Visit Data Across Multiple Sources

By Shaked Markovich
Nov 2024

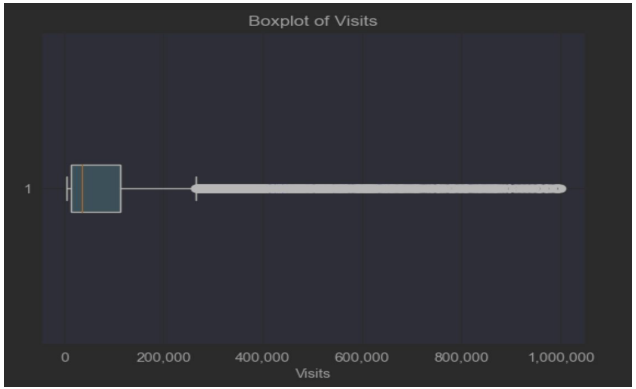
Objective

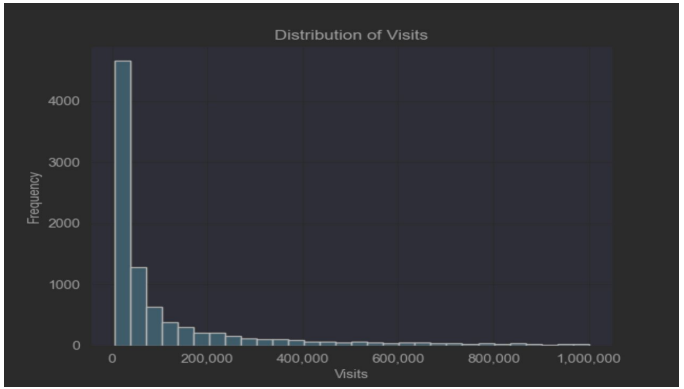
Objective:

- Compare three sources of site visit data against a learning set.
- Explore ways to align the scales between sources and learning data.
- Apply various statistical methods to evaluate performance (e.g., regression, KS test, PCA).

Exploratory Data Analysis (EDA) - Distribution

- **Key Observations:**
  - Right-skewed distributions in visit counts.
  - Significant concentration of sites with small visit counts.
  - Visualization: Histograms and boxplots showing distribution across sources.
- **Insight:** This aligns with the "80/20 Rule" — a small number of sites drive most of the visits.





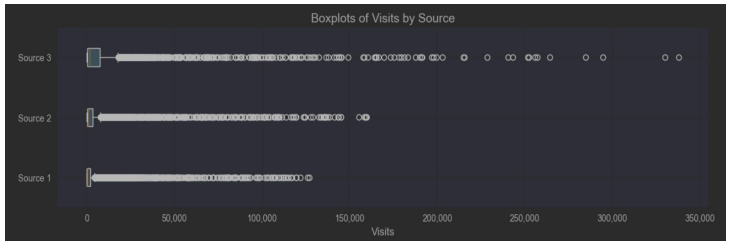
EDA - Scale Differences Across Sources

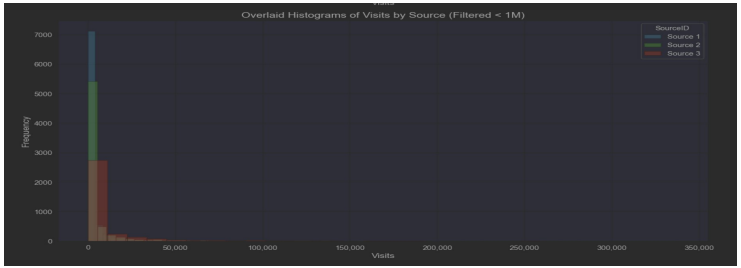
Key Observations:

- Basic statistics (mean, median, max, min) for each source.
- Histogram, boxplot, and CDF show varying distributions across sources.

Insight: Differences may be due to:

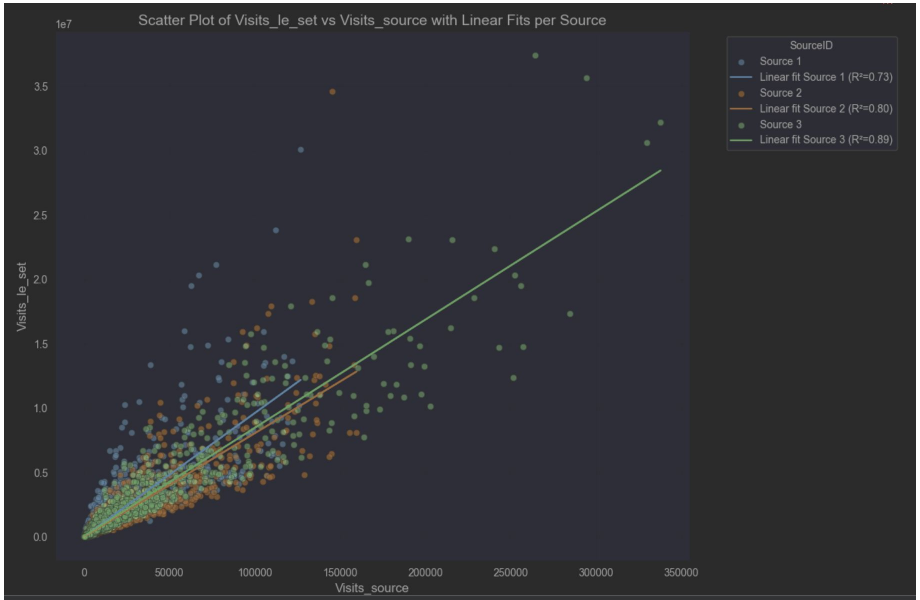
- Diverse underlying distributions.
- Different demographic groups or time periods of data collection.





Correlation Analysis

Insight: Observing the relationships and identifying potential outliers




```
R² value for Source 1: 0.7289
R² value for Source 2: 0.7964
R² value for Source 3: 0.8947
```

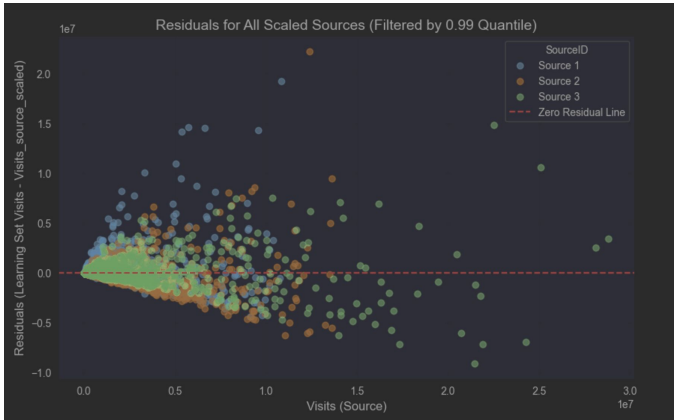
Scaling and Model Fitting

Goal: Adjust scales using models to reduce discrepancies.

Methods:

- Linear Regression used to scale `Visits_source` to match the learning set.
- RMSE, MAE, MAPE, and MSE used to evaluate model performance.

	RMSE	MAPE	MAE	MSE
1	654915	127	172440	428913471416
2	636080	54	170559	404597793902
3	835340	149	287837	697793692482



key findings

Source 1: Highest overlap with the learning set, but low R^2.

Source 3: Least overlap with the learning set, but high R^2 and significant KS difference.

Source 2: Best performing across error metrics (lowest RMSE, MAE, MSE).

Model Comparison: Baseline vs. Ensemble

Ensemble Learning Approach:

- Combined predictions from multiple models (XGBoost) using a Voting Regressor.
- PCA used to combine features from all sources.

Results:

- Ensemble did not significantly reduce error.
- Using all sources as features reduced error only when compared to Source 2 alone.
- PCA combined sources most effectively, leading to the lowest error.

```
RMSE for model using Visits_source_scaled_1: 811346.4059
RMSE for model using Visits_source_scaled_2: 695047.0648
RMSE for model using Visits_source_scaled_3: 549590.2494
```

```
RMSE for PCA-based model: 510466.4756
```

Conclusion and Key Insights

Key Insights:

- **Source 1**: Strong overlap with learning set, but poor predictive power (low R^2).
- **Source 3**: Minimal overlap, but better predictive power (high R^2), though statistically different from the learning set.
- **Source 2**: The most reliable predictor with the lowest error across all metrics.
- **Ensemble and PCA**: Combining scaled sources and using PCA resulted in better performance than using individual sources.

Next Steps:

- Further tuning of ensemble models.
- Explore advanced feature engineering for better prediction accuracy.