**Customer Transaction Process**

**Prediction Project Report**

TEB Arf Case Study

**Eser Inan Arslan**

**April 2024**

**TEB-ARF Data Science Challenge**

## 

## **Algorithm Evaluation Summary**

### **Algorithms Tested:**

#### LSTM

#### ResNet

#### GNN (Graph Neural Network)

### **Additional Model Considerations:**

#### Experimenting with Graph Embedding for feature derivation or direct modeling with GNN would provide additional insights.

### **Success Criteria:**

## The candidate is expected to choose appropriate metrics for the classification problem.

## **Performance Metrics**

#### **LSTM Model (Threshold: 0.5)**

* Accuracy**:** 91.57%
* Classification Report:

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
|  | 0.0 | 0.69 | 0.46 | 2340 |
|  | 1.0 | 0.98 | 0.95 | 20056 |
| accuracy |  |  | 0.92 | 22396 |
| macro avg | 0.81 | 0.66 | 0.71 | 22396 |
| weighted avg | 0.90 | 0.92 | 0.90 | 22396 |

#### **ResNet Model (Threshold: 0.5)**

* Accuracy: 89.63%
* Classification Report:

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
|  | 0.0 | 0.50 | 0.43 | 2340 |
|  | 1.0 | 0.96 | 0.94 | 20056 |
| accuracy |  |  | 0.90 | 22396 |
| macro avg | 0.72 | 0.67 | 0.69 | 22396 |
| weighted avg | 0.89 | 0.90 | 0.89 | 22396 |

#### **GNN Model (Threshold: 0.75)**

* Accuracy: 89.55%
* Classification Report:

|  | precision | recall | f1-score | support |
| --- | --- | --- | --- | --- |
|  | 0.0 | 0.00 | 0.00 | 2340 |
|  | 1.0 | 1.00 | 0.94 | 20056 |
| accuracy |  |  | 0.90 | 22396 |
| macro avg | 0.45 | 0.50 | 0.47 | 22396 |
| weighted avg | 0.80 | 0.90 | 0.85 | 22396 |

## **Conclusion:**

Detailed Positive Case Counts:

* Total Positive Case Number : 20056
* LSTM True Positive Case Number : 19696
* ResNet True Positive Case Number : 19187
* GNN True Positive Case Number : 17545

This summary provides an overview of the evaluated algorithms (LSTM, ResNet and GNN) and highlights the performance metrics achieved. Additionally, it outlines the importance of considering Graph Neural Networks (GNN) for potential future model enhancements.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Summary of Process**

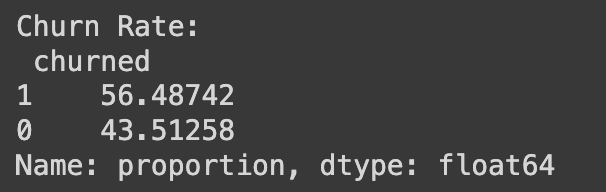
### **EDA**

In the exploratory data analysis (EDA) phase of the Customer Lifetime Value (CLV) prediction project for Corpay Company, several key observations were made. Firstly, the shape of the dataset is (101350, 104), indicating a considerable amount of data with numerous features for analysis.

During the data preprocessing stage, it was noted that the dataset spans from August 13, 1990, to March 15, 2024, with 'start\_date' serving as the beginning point and 'last\_transaction\_date' as the closing threshold. This extensive timeline prompted a closer examination of more recent data. Notably, observations were made considering two significant dates in the UK context: January 31, 2020, marking the first Covid-19 case in the UK, and June 23, 2016, the date of the UK's withdrawal from the EU, commonly known as Brexit. Truncating the dataset based on these dates revealed insights into the impact of major events on the data volume.

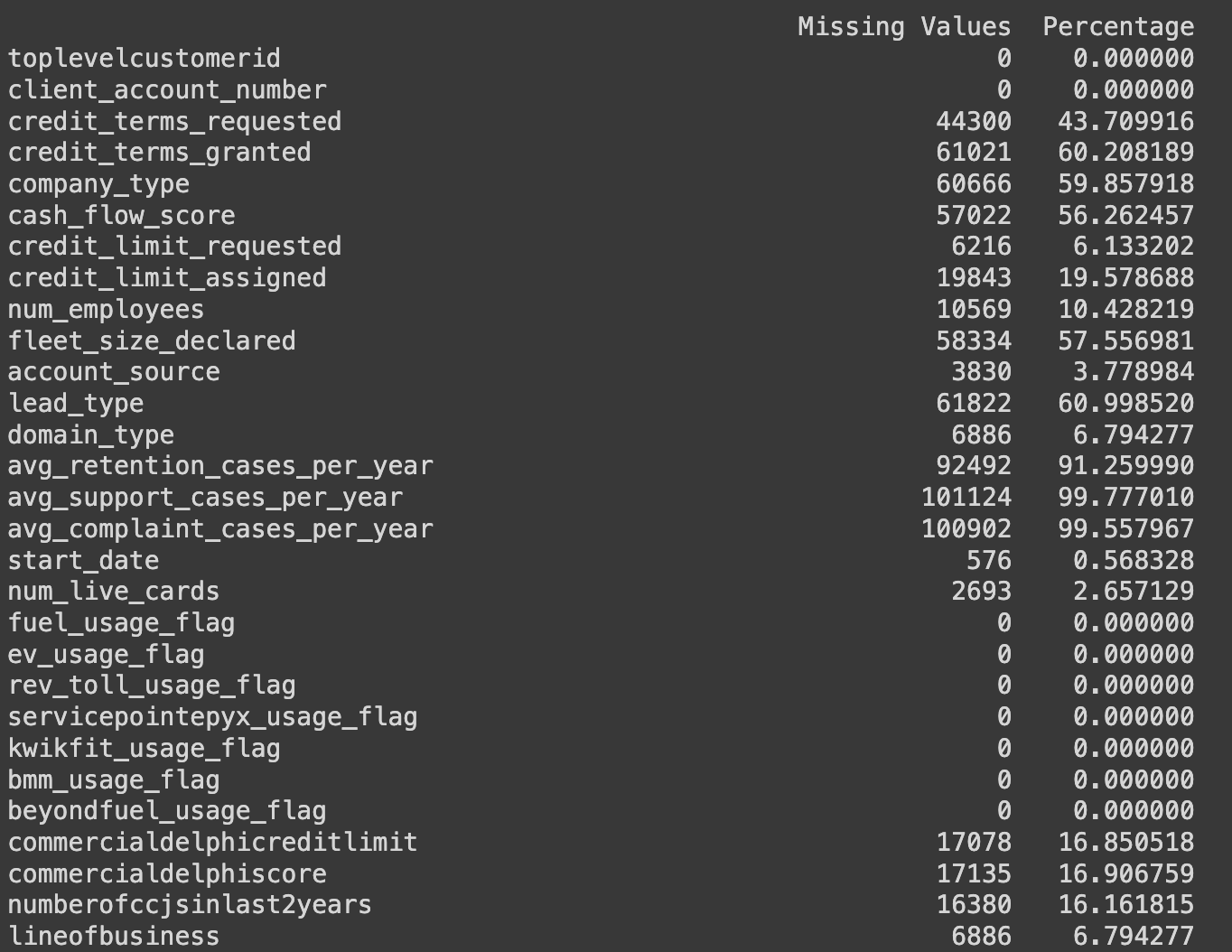
Concerning the features suitable for CLV prediction, 'total\_spend\_monthly\_avg', 'total\_transactions\_monthly\_avg', and 'total\_rev\_monthly\_avg' exhibited high potential due to their similarities and relevance to the CLV problem.

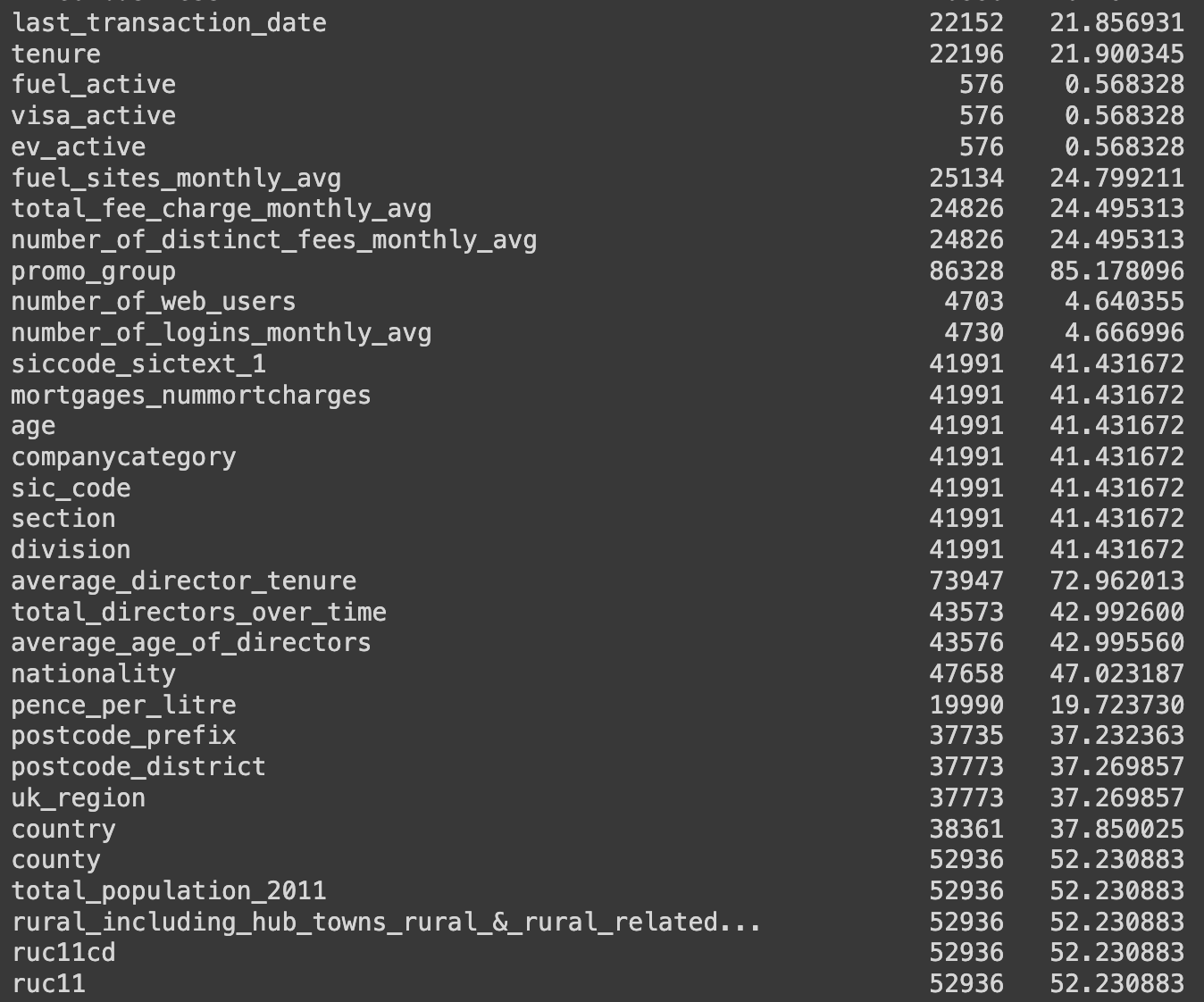
Furthermore, the churn analysis revealed that 56% of the total users, amounting to 57250 individuals, had churned. The last transaction date of the most recent churned customer was recorded as '2023-12-17', highlighting recent churn events.

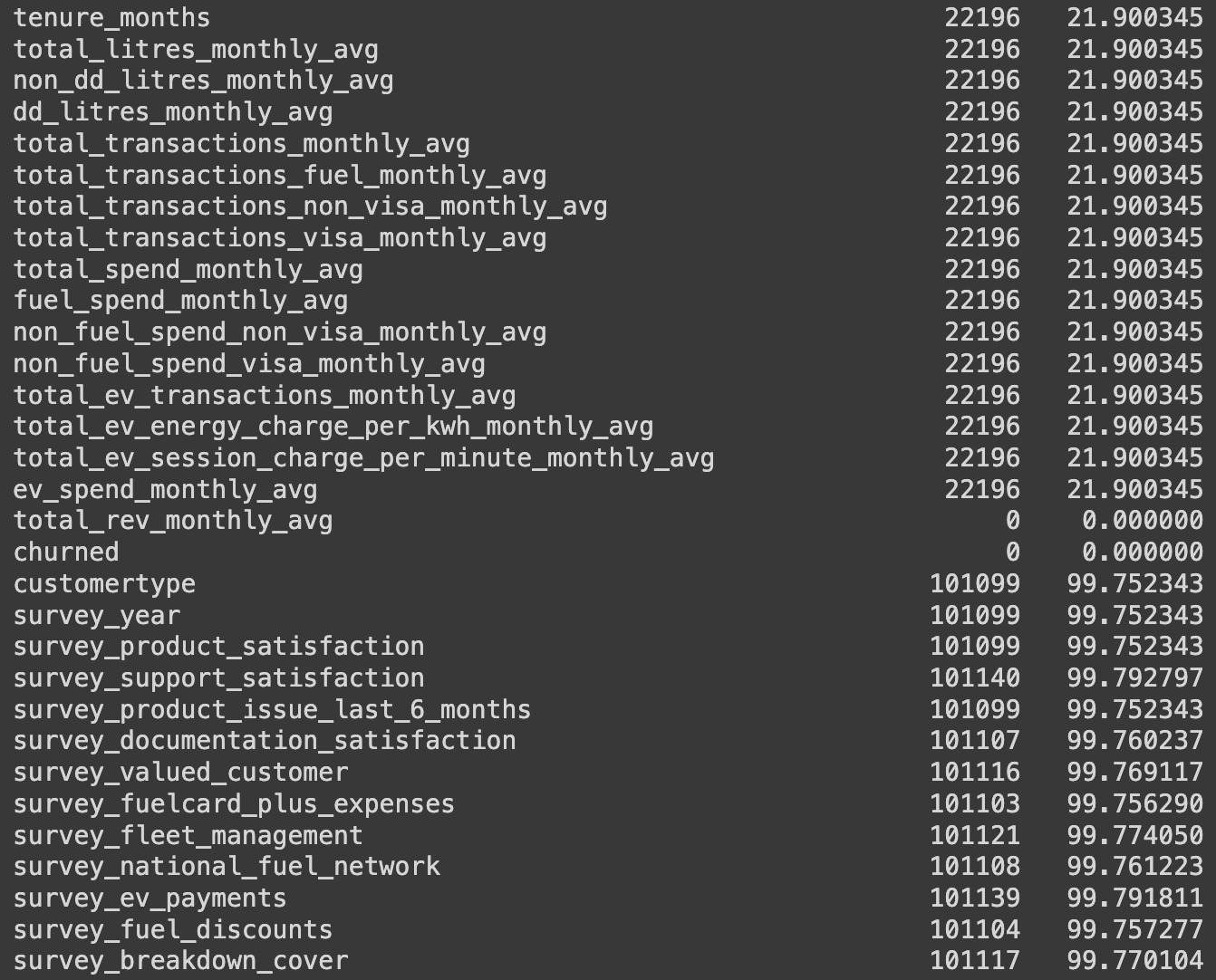


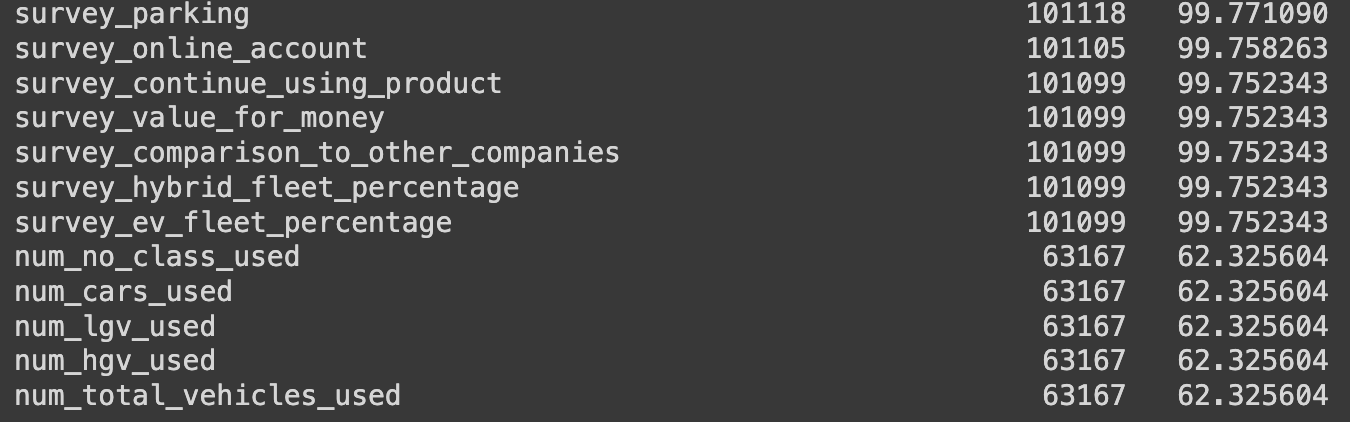
To ensure the accuracy of CLV prediction based on 'total\_rev\_monthly\_avg', data integrity was crucial. Accordingly, records with negative values in 'total\_rev\_monthly\_avg' were excluded from the training set. Moreover, descriptive statistics revealed insights into the distribution of 'total\_rev\_monthly\_avg', with an average of 60.5, a standard deviation of 298, and a maximum value of 34584.

Missing data analysis revealed varying degrees of data completeness across features, with some columns exhibiting more than 60% missing values. Features with substantial missing data were dropped to maintain the integrity of the analysis.

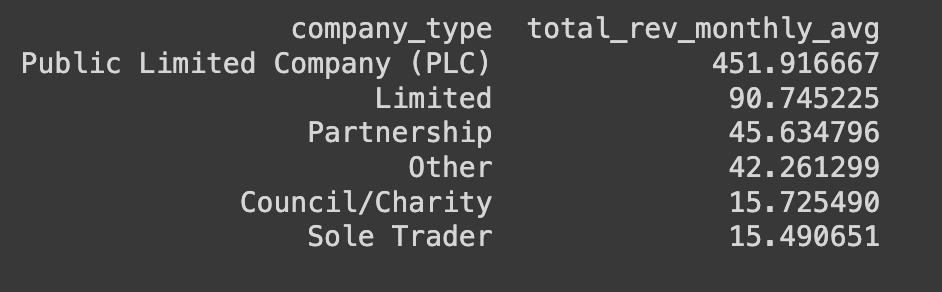




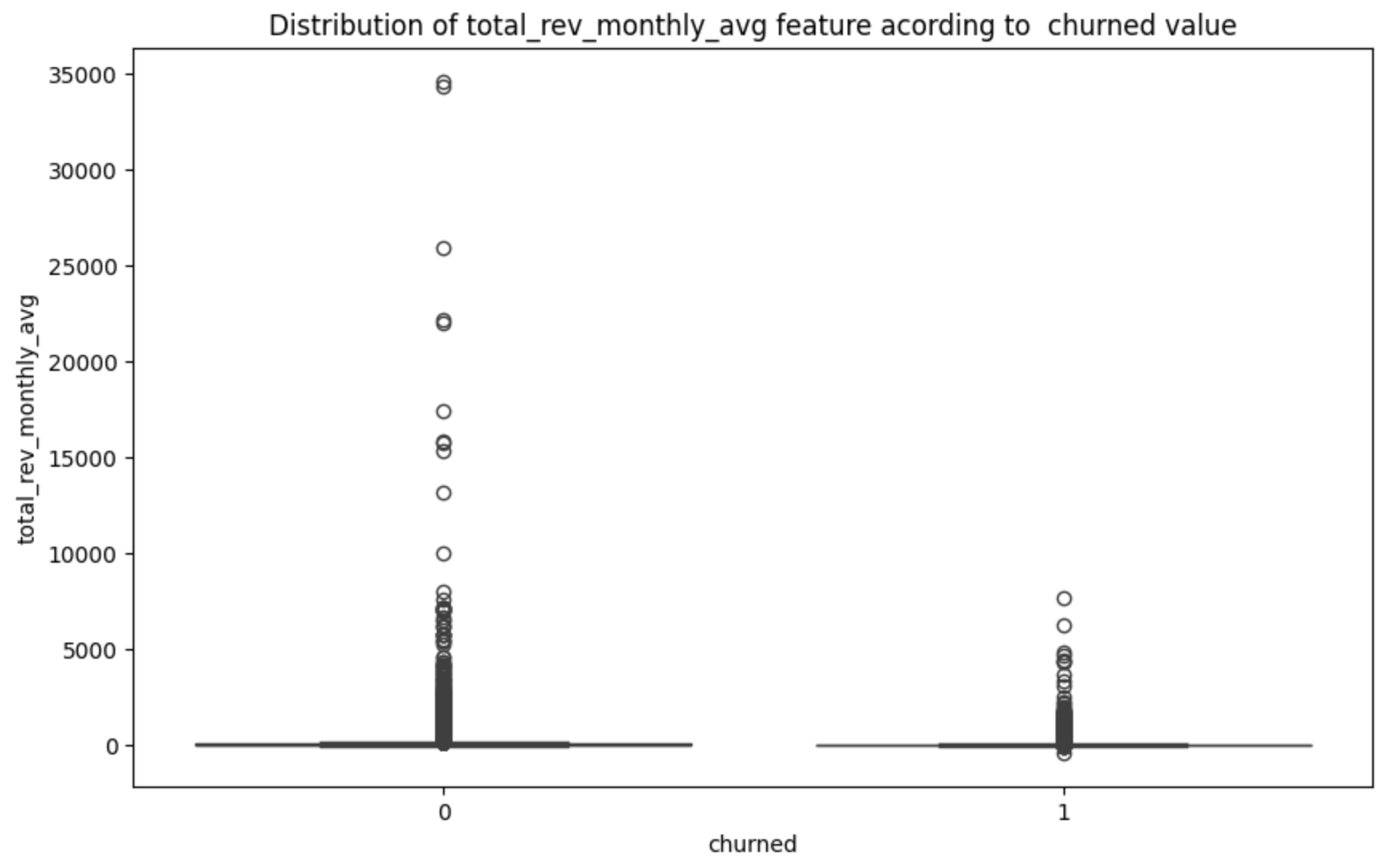




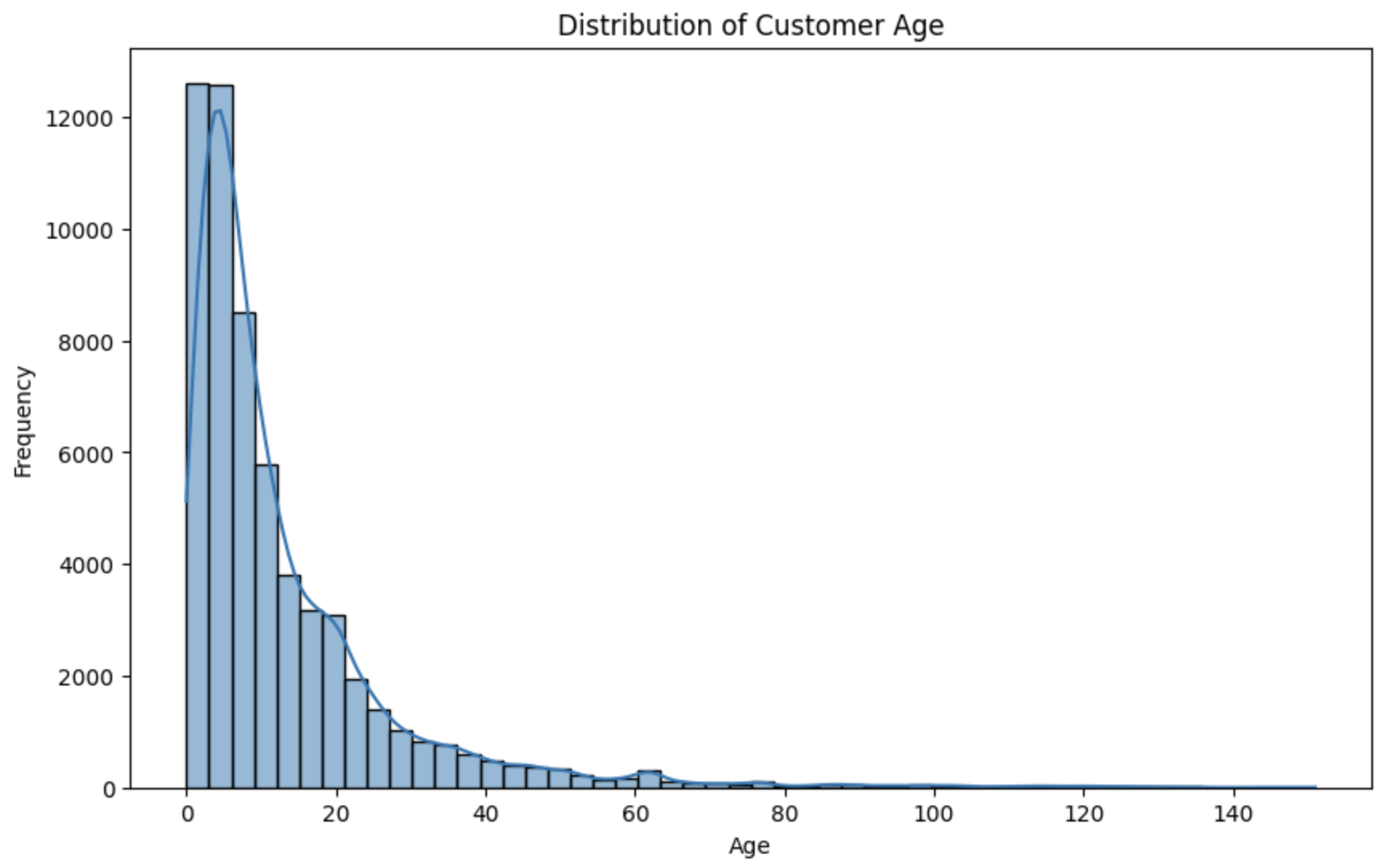
The distribution of 'total\_rev\_monthly\_avg' by company types shed light on spending patterns, with Public Limited Companies (PLCs) leading with an average of 451. Limited companies followed with an average of 90, while Council/Charity and Sole Traders recorded the lowest averages at 15.



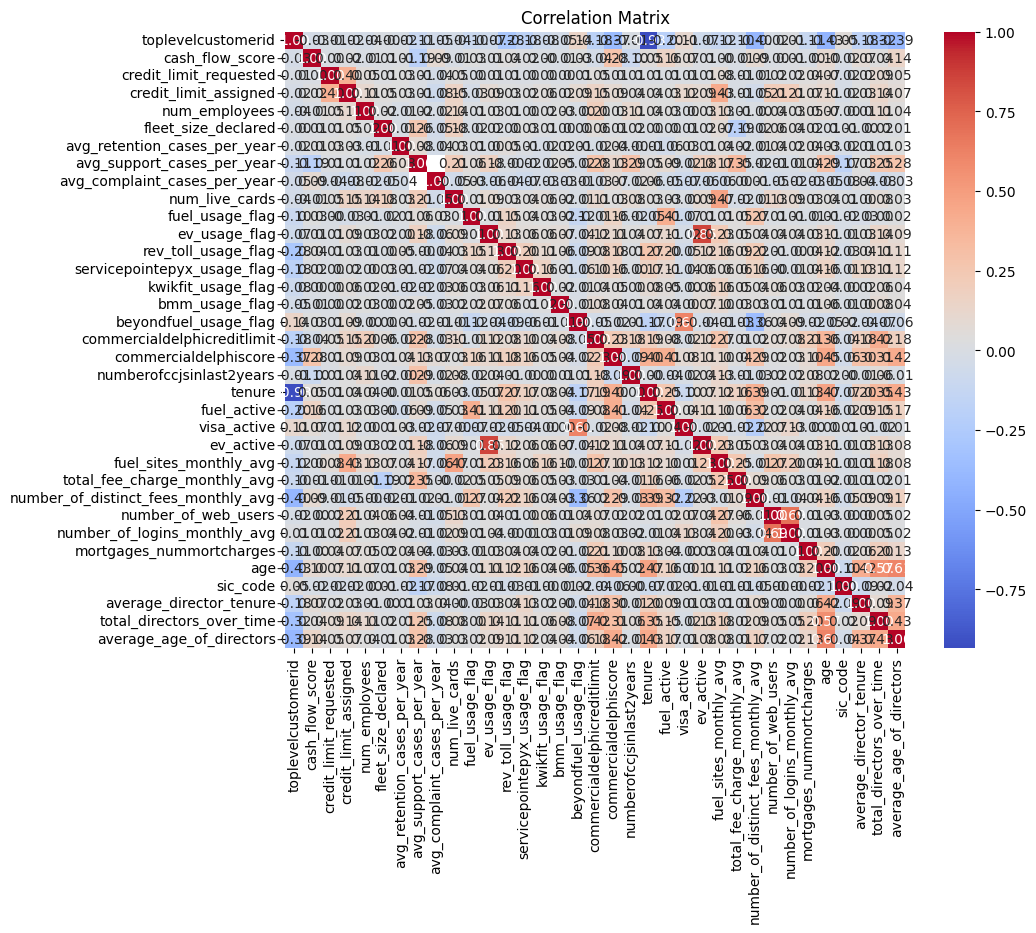
Examining churned versus non-churned customers' 'total\_rev\_monthly\_avg' averages provided insights into spending behaviors, with churned customers averaging around 29, compared to 101 for non-churned customers.



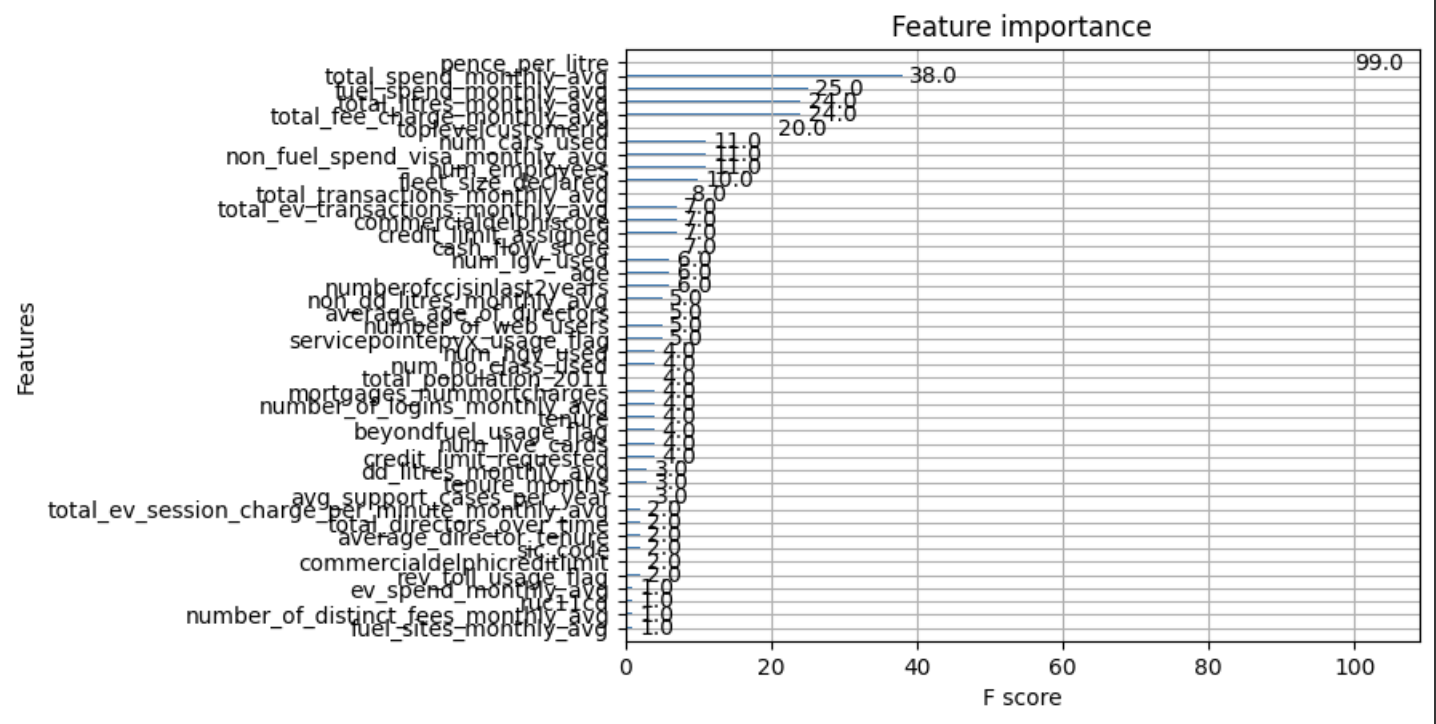
Analysis of the 'Age' feature revealed a skewed distribution, with a significant portion of users clustered below the age of 20.



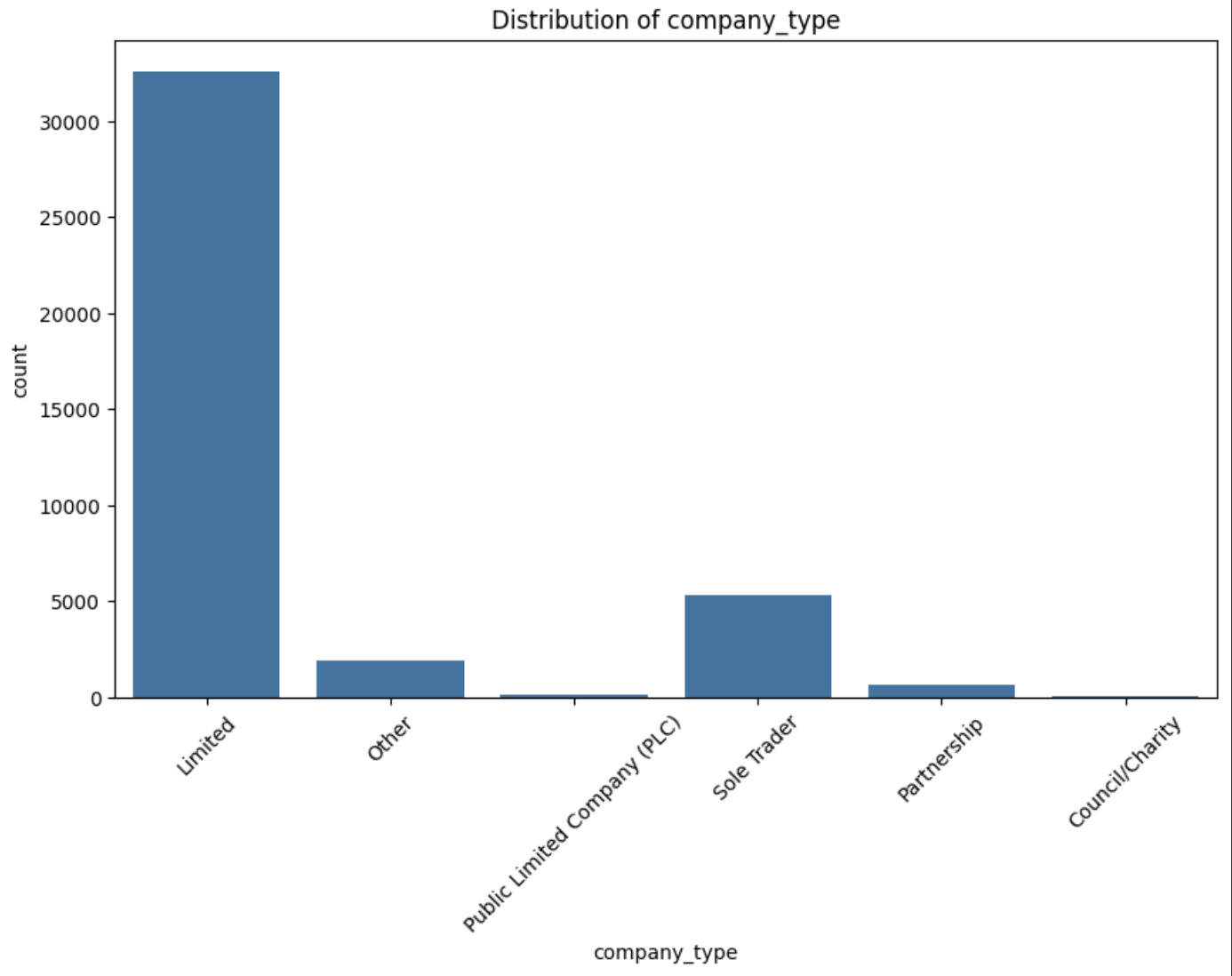
Correlation analysis identified 15 features with strong correlations (above 0.5) with 'total\_rev\_monthly\_avg', providing valuable insights into potential predictors of revenue.

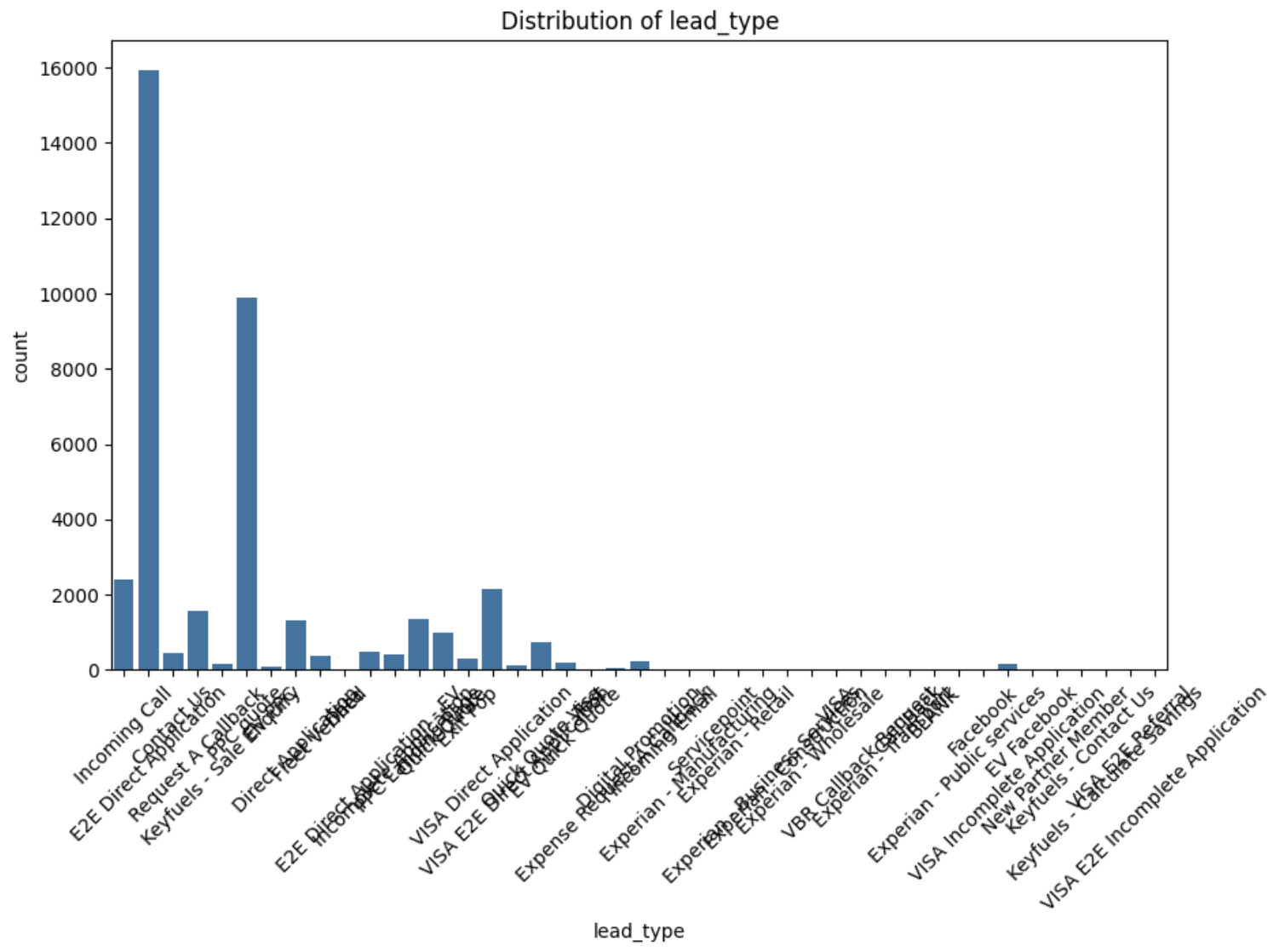


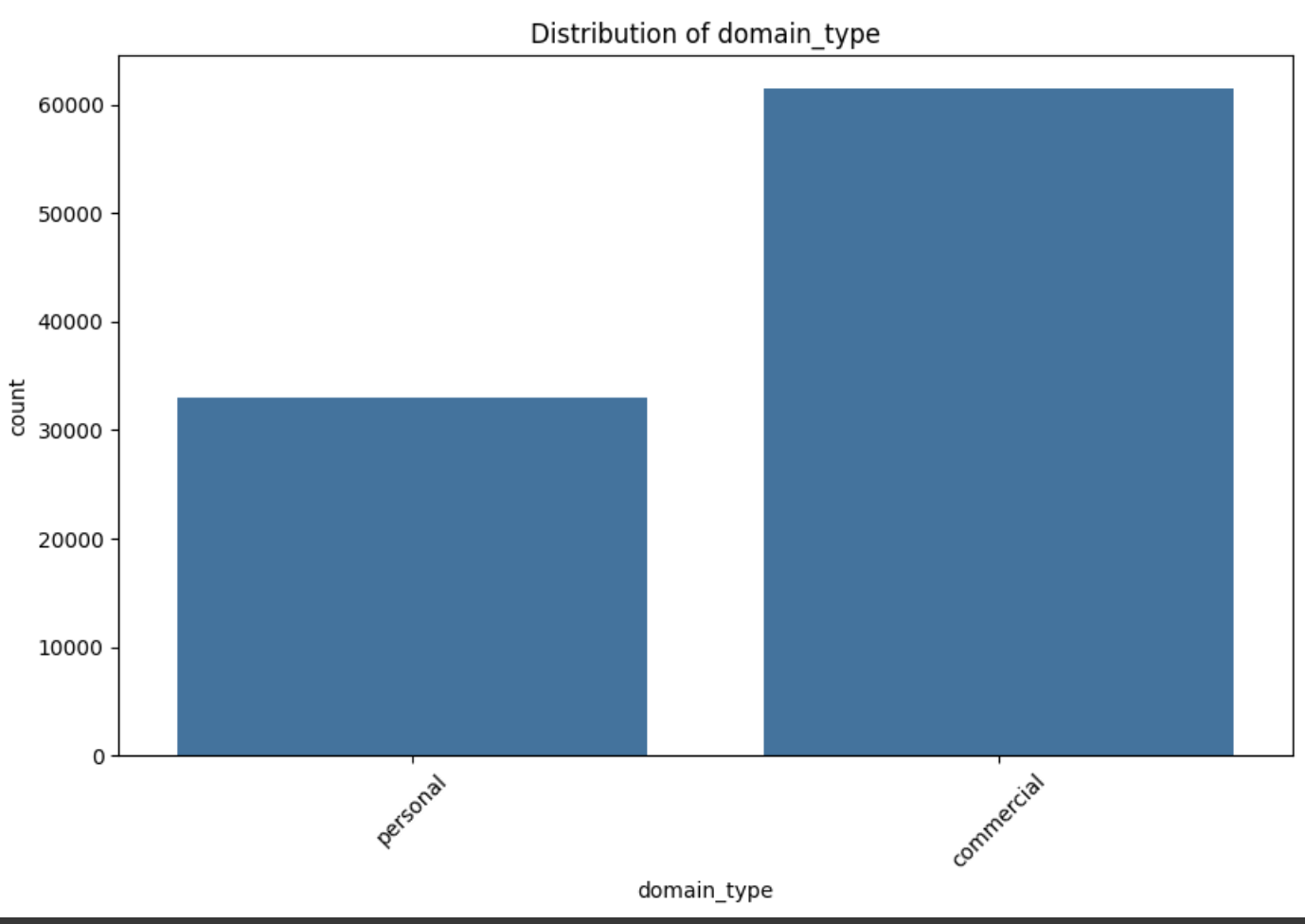
Additionally, feature importance evaluation using the XG Boost method yielded results consistent with correlation findings.



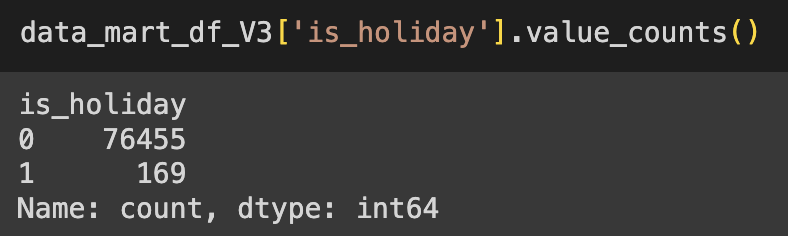
Moreover, distribution analysis based on 'company\_type', 'lead\_type', and 'domain\_type' highlighted notable trends in company characteristics.







Lastly, holiday analysis revealed that 169 'start\_date' values coincided with official holidays in the UK, providing additional context for the dataset.



Overall, the EDA phase provided valuable insights into the dataset's characteristics, distribution, and potential predictors, laying a solid foundation for subsequent analysis and modeling tasks.

### **PREPROCESS**

To begin with, I implemented a custom function aimed at reducing the memory usage of the dataset, effectively mitigating memory-related issues. This function was applied multiple times, and upon the initial execution, it resulted in a substantial reduction of memory usage, from 37.89 Mb, representing a 52.9% decrease.

Next, I focused on logical mappings and data deduplication for fields such as company type. Additionally, to normalize columns containing lengthy textual values, I employed Natural Language Toolkit (NLTK) for tasks like tokenization, stopword removal, and regex tokenization, ensuring consistency across the dataset.

Identifying data errors was imperative, prompting the removal of records with negative values in fields like 'total\_spend\_monthly\_avg', 'total\_transactions\_monthly\_avg', and 'total\_rev\_monthly\_avg'. This cleanup operation resulted in the removal of approximately 22,000 erroneous records.

Given the dataset's extensive temporal range spanning about 34 years, I opted to establish a threshold date of '23/06/2016' (Brexit date). Subsequently, only data post this threshold date was deemed relevant for model training, disregarding potentially outdated information.

Aiming to enrich the model's feature set from an engineering perspective, I integrated UK holiday data and flagged holidays within the dataset, providing the model with additional contextual information.

Calculating the duration of customer activity by computing the difference between 'last\_transaction\_date' and 'start\_date' yielded the 'customer\_activity\_duration' feature, contributing to a more comprehensive understanding of customer behavior.

Following this, date columns were updated based on the difference in days relative to the maximum value of 'last\_transaction\_date', ensuring temporal accuracy and relevance in the dataset.

I then computed the 'transaction\_diversity' metric based on the presence of flags in features such as 'fuel\_usage\_flag', 'ev\_usage\_flag', and 'rev\_toll\_usage\_flag', capturing the diversity of customer transaction behaviors.

Custom encoding schemes were devised for specific data attributes like 'credit\_terms\_requested', optimizing the representation of categorical data.

For handling missing values, 'Other' was imputed for object types, while the mean of respective features was utilized for numeric data, ensuring data completeness and integrity.

Categorical features underwent encoding based on their diversity levels, with LabelEncoder used for those with high diversity and one-hot encoding (get\_dummies) employed for features with lower diversity, enhancing the model's ability to interpret categorical data accurately.

Normalization of numeric columns was executed via four distinct methodologies, including binning, MinMaxScaler, percentile-based grouping, and manual threshold-based grouping, ensuring consistency and comparability across numerical features.

The dataset was partitioned into training and testing sets based on a temporal criterion, with the last 7 days of data reserved for prediction and preceding data utilized for model training, leveraging a threshold date of '2024-03-09'.

Lastly, the target variable 'total\_rev\_monthly\_avg' was designated for prediction, and the resulting dataset shapes were recorded as follows:

X\_train.shape (43192, 117),   
y\_train.shape (43192,),   
X\_test.shape (11739, 117),   
y\_test.shape (11739,).

### **Model Selection**

For the predictive modeling task, I opted to explore two distinct types of algorithms to ensure a comprehensive comparison. Considering factors such as dataset size, problem type, data diversity, and computational resources, I selected one classical machine learning algorithm and one deep learning algorithm.

#### **Classical Machine Learning Algorithm: XGBoost Regressor**

The XGBoost Regressor was chosen due to its versatility and robust performance in various regression tasks. It's particularly well-suited for handling large datasets with high dimensionality.

#### **Deep Learning Algorithm: Recurrent Neural Network (RNN)**

Given the sequential nature of the data and the potential for capturing temporal dependencies, I decided to employ a Recurrent Neural Network (RNN). RNNs are adept at modeling sequential data and have shown considerable success in time series forecasting tasks.

#### **Parameter Tuning**

To optimize the performance of each model, I utilized the grid search method for hyperparameter tuning. Grid search allows for an exhaustive search over a predefined hyperparameter space, enabling the identification of the optimal combination of parameters.

#### **Error Measurement**

For evaluating the performance of the models, two key error metrics were employed:

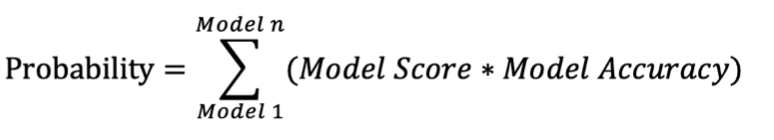
1. **Mean Absolute Percentage Error (MAPE):** MAPE measures the accuracy of the models in predicting continuous variables, providing insight into the average percentage deviation between predicted and actual values.
2. **Root Mean Squared Error (RMSE):** RMSE quantifies the average magnitude of the errors between predicted and actual values, providing a measure of the model's predictive accuracy.

#### **Model Conclusion**

The utilization of XGBoost Regressor and Recurrent Neural Network models, coupled with thorough parameter tuning using grid search, allowed for a comprehensive exploration of predictive modeling techniques. The evaluation of models using MAPE and RMSE will provide valuable insights into their performance and suitability for the given task.

#### **Churn Risk**

Firstly, I delved into the issue of churn prediction, focusing on recent customer behavior within the last 6 months. Utilizing machine learning algorithms such as KNN, Random Forest, and Naive Bayes, I computed churn probabilities and weighted them based on model accuracy to derive a comprehensive churn score.



The results yielded the following insights: out of a total of 3151 records, 2974 were identified as negative records, indicating non-churn instances, while 177 records were classified as positive, signifying churn cases. Among these, 2599 records were accurately classified as true negatives, and 177 as true positives, demonstrating the model's efficacy in distinguishing between churn and non-churn instances.

Of particular note is the subset of 375 customers identified as having a high likelihood of churn. This subset represents a critical segment that requires immediate attention and intervention to prevent churn. Given that these customers contribute to a total monthly revenue of $4276, their retention is paramount to preserving revenue streams and sustaining business growth.

In conclusion, the churn prediction analysis underscores the importance of proactive churn management strategies, emphasizing the need to identify and engage with at-risk customers effectively. By leveraging insights from predictive modeling, businesses can implement targeted retention efforts aimed at mitigating churn and fostering long-term customer loyalty.

#### **Customer Groups**

##### Churned Customers with Below-Average Monthly Revenue (Cluster 1):

* This cluster comprises a substantial number of customers, totaling 50,190, who have churned and consistently generated below-average monthly revenue.
* These customers may represent a segment that faced challenges or dissatisfaction with the service/product offering, leading to their decision to churn.
* Potential Actions: Targeted re-engagement campaigns could be implemented to understand the reasons behind their dissatisfaction and offer tailored solutions to encourage them to return. Additionally, offering incentives or discounts may help re-attract these customers.

##### Churned Customers with Above-Average Monthly Revenue (Cluster 2):

* This cluster consists of 7,060 customers who churned despite having above-average monthly revenue.
* These customers might have churned due to factors unrelated to financial considerations, such as poor customer service experiences, changing preferences, or external factors.
* Potential Actions: Conducting exit surveys or interviews with these customers can provide valuable insights into the specific reasons for churn. Strategies to improve customer experience, enhance product/service offerings, or address pain points identified through feedback can be implemented to reduce churn in this segment.

##### Non-Churned Customers with Below-Average Monthly Revenue (Cluster 3):

* This cluster includes 28,618 customers who have not churned and consistently maintained below-average monthly revenue.
* Despite their lower spending, these customers have remained loyal to the business, indicating satisfaction with the product/service or other non-monetary benefits they receive.
* Potential Actions: Implementing loyalty programs, personalized promotions, or upselling/cross-selling initiatives can encourage these customers to increase their spending and further solidify their loyalty.

##### Non-Churned Customers with Above-Average Monthly Revenue (Cluster 4):

* The final cluster comprises 15,482 customers who have not churned and consistently generated above-average monthly revenue.
* These customers represent the most valuable segment, contributing significantly to the business's revenue stream while demonstrating loyalty and satisfaction with the offerings.
* Potential Actions: Providing exclusive benefits, VIP treatment, or premium services to acknowledge and reward their loyalty can further strengthen the relationship with these high-value customers and encourage long-term retention.