Customer Lifetime Value Project Report

Eser Inan Arslan
April 2024

Corpay Data Science Challenge

Company Overview:

Corpay is a leading firm in corporate payment solutions. We are excited to assess your skills and potential contributions to our data-driven culture through this challenge.

Challenge Objective:

You are tasked with analyzing a dataset provided by us to uncover insights and drive strategic business outcomes. This is a real-world example of a data science project; as such, you are provided a dataset and some documentation, and both may be incomplete or occasionally inaccurate. You are welcome to ask any question on data, documentation or business needs.

Tasks:

1. Data Cleaning and Preparation:

- Handle missing values and anomalies.
- Engineer new features that could be helpful for the analysis.

2. Exploratory Data Analysis (EDA):

- Provide a thorough statistical summary of the components of the dataset.
- Visualize important relationships between key variables.

3. Predictive Modeling:

- Develop a model to predict Customer Lifetime Value
- Evaluate the performance of your model using appropriate metrics.

4. Segment Customers:

- Create a small number of customer segments
- Describe those segments and potential actions the business could take to increase value from each segment.

5. Presentation of Findings:

- Prepare a report summarizing your methodologies, findings, performance.
- Include key insights from the data that can aid in decision-making.

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.

```
Churn Rate:
 churned
1    56.48742
0    43.51258
Name: proportion, dtype: float64
```

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.

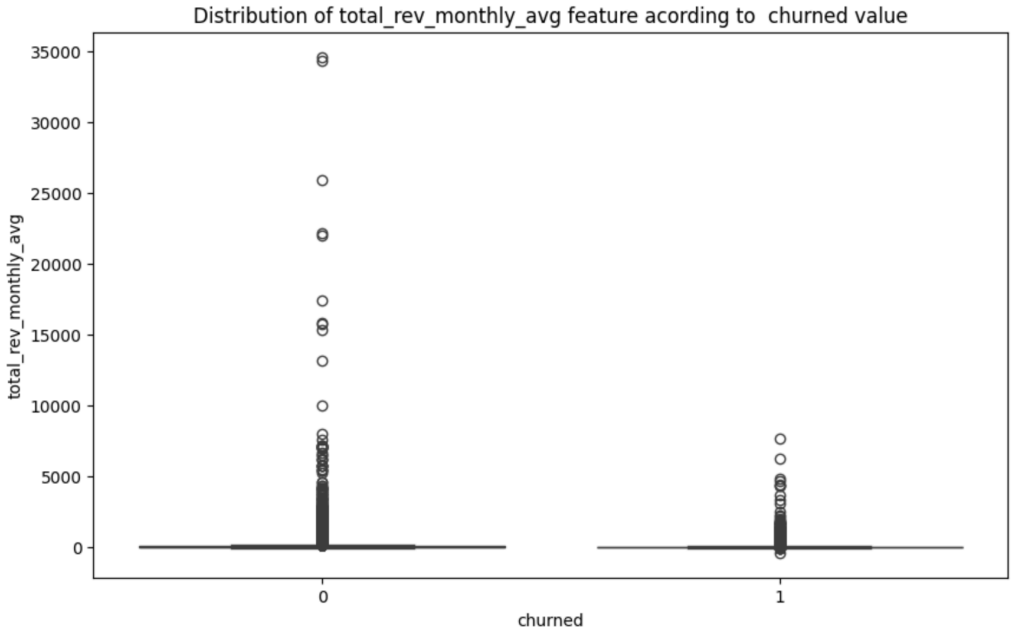
	Missing Values	Percentage
toplevelcustomerid	0	0.000000
client_account_number	0	0.000000
credit_terms_requested	44300	43.709916
credit_terms_granted	61021	60.208189
company_type	60666	59.857918
cash_flow_score	57022	56.262457
credit_limit_requested	6216	6.133202
credit_limit_assigned	19843	19.578688
num_employees	10569	10.428219
fleet_size_declared	58334	57.556981
account_source	3830	3.778984
lead_type	61822	60.998520
domain_type	6886	6.794277
avg_retention_cases_per_year	92492	91.259990
avg_support_cases_per_year	101124	99.777010
avg_complaint_cases_per_year	100902	99.557967
start_date	576	0.568328
num_live_cards	2693	2.657129
fuel_usage_flag	0	0.000000
ev_usage_flag	0	0.000000
rev_toll_usage_flag	0	0.000000
servicepointepyx_usage_flag	0	0.000000
kwikfit_usage_flag	0	0.000000
bmm_usage_flag	0	0.000000
beyondfuel_usage_flag	0	0.000000
commercialdelphicreditlimit	17078	16.850518
commercialdelphiscore	17135	16.906759
numberofccjsinlast2years	16380	16.161815
lineofbusiness	6886	6.794277
last_transaction_date	22152	21.856931
tenure	22196	21.900345
fuel_active	576	0.568328
visa_active	576	0.568328
ev_active	576	0.568328
fuel_sites_monthly_avg	25134	24.799211
total_fee_charge_monthly_avg	24826	24.495313
number_of_distinct_fees_monthly_avg	24826	24.495313
promo_group	86328	85.178096
number_of_web_users	4703	4.640355
number_of_logins_monthly_avg	4730	4.666996
siccode_sictext_1	41991	41.431672
mortgages_nummortcharges	41991	41.431672
age	41991	41.431672
companycategory	41991	41.431672
sic_code	41991	41.431672
section	41991	41.431672
division	41991	41.431672
average_director_tenure	73947	72.962013
total_directors_over_time	43573	42.992600
average_age_of_directors	43576	42.995560
nationality	47658	47.023187
pence_per_litre	19990	19.723730
postcode_prefix	37735	37.232363
postcode_district	37773	37.269857
uk_region	37773	37.269857
country	38361	37.850025
county	52936	52.230883
total_population_2011	52936	52.230883
rural_including_hub_towns_rural_&_rural_related...	52936	52.230883
ruc11cd	52936	52.230883
ruc11	52936	52.230883

tenure_months	22196	21.900345
total_litres_monthly_avg	22196	21.900345
non_dd_litres_monthly_avg	22196	21.900345
dd_litres_monthly_avg	22196	21.900345
total_transactions_monthly_avg	22196	21.900345
total_transactions_fuel_monthly_avg	22196	21.900345
total_transactions_non_visa_monthly_avg	22196	21.900345
total_transactions_visa_monthly_avg	22196	21.900345
total_spend_monthly_avg	22196	21.900345
fuel_spend_monthly_avg	22196	21.900345
non_fuel_spend_non_visa_monthly_avg	22196	21.900345
non_fuel_spend_visa_monthly_avg	22196	21.900345
total_ev_transactions_monthly_avg	22196	21.900345
total_ev_energy_charge_per_kwh_monthly_avg	22196	21.900345
total_ev_session_charge_per_minute_monthly_avg	22196	21.900345
ev_spend_monthly_avg	22196	21.900345
total_rev_monthly_avg	0	0.000000
churned	0	0.000000
customertype	101099	99.752343
survey_year	101099	99.752343
survey_product_satisfaction	101099	99.752343
survey_support_satisfaction	101140	99.792797
survey_product_issue_last_6_months	101099	99.752343
survey_documentation_satisfaction	101107	99.760237
survey_valued_customer	101116	99.769117
survey_fuelcard_plus_expenses	101103	99.756290
survey_fleet_management	101121	99.774050
survey_national_fuel_network	101108	99.761223
survey_ev_payments	101139	99.791811
survey_fuel_discounts	101104	99.757277
survey_breakdown_cover	101117	99.770104
survey_parking	101118	99.771090
survey_online_account	101105	99.758263
survey_continue_using_product	101099	99.752343
survey_value_for_money	101099	99.752343
survey_comparison_to_other_companies	101099	99.752343
survey_hybrid_fleet_percentage	101099	99.752343
survey_ev_fleet_percentage	101099	99.752343
num_no_class_used	63167	62.325604
num_cars_used	63167	62.325604
num_lgv_used	63167	62.325604
num_hgv_used	63167	62.325604
num_total_vehicles_used	63167	62.325604

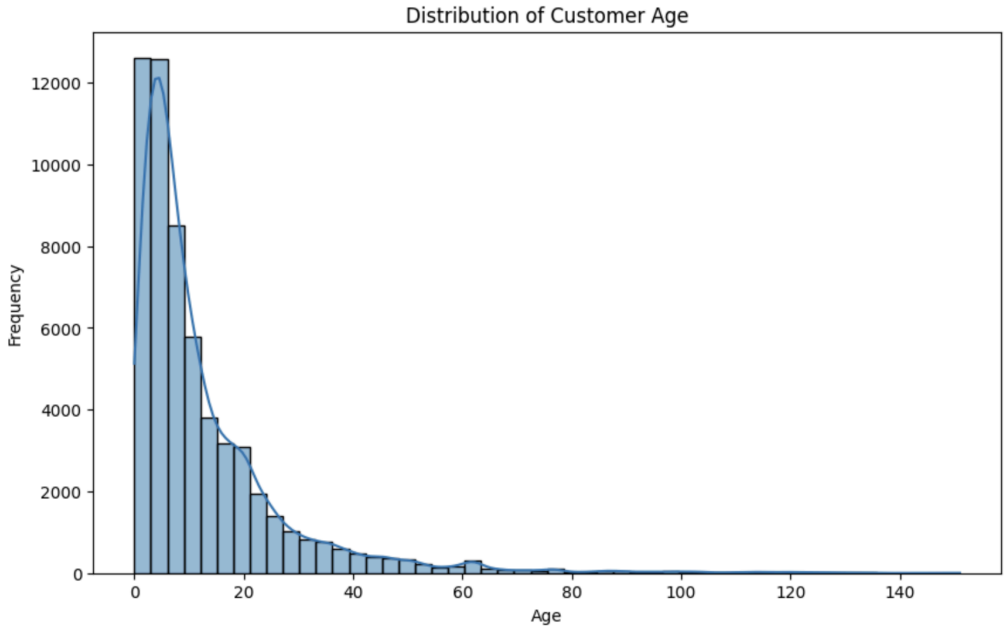
The distribution of 'total_rev_monthly_avg' by company types shed light on spending patterns, with Public Limited Companies (PLCs) leading with an average of 451. Limited companies followed with an average of 90, while Council/Charity and Sole Traders recorded the lowest averages at 15.

company_type	total_rev_monthly_avg
Public Limited Company (PLC)	451.916667
Limited	90.745225
Partnership	45.634796
Other	42.261299
Council/Charity	15.725490
Sole Trader	15.490651

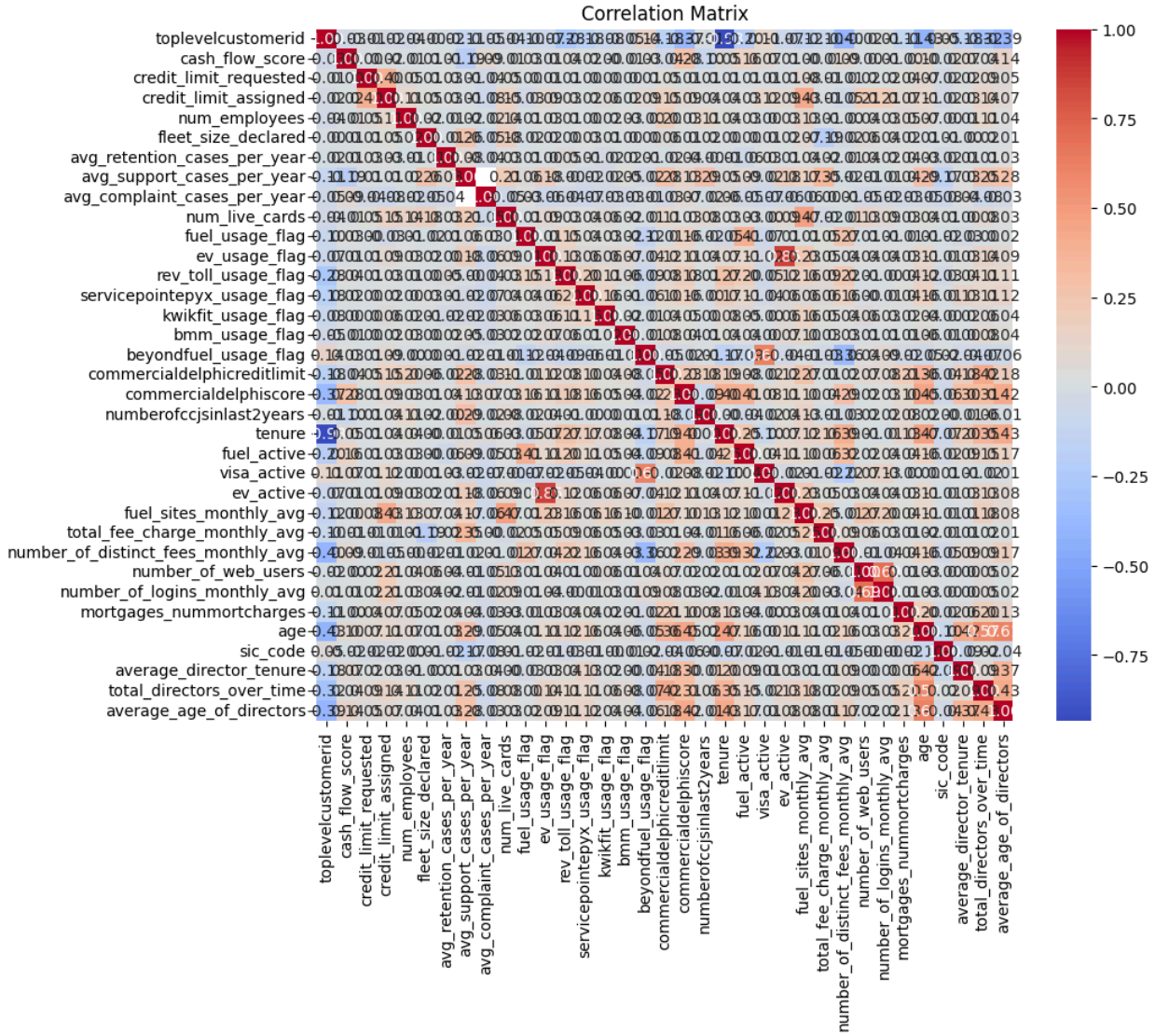
Examining churned versus non-churned customers' 'total_rev_monthly_avg' averages provided insights into spending behaviors, with churned customers averaging around 29, compared to 101 for non-churned customers.



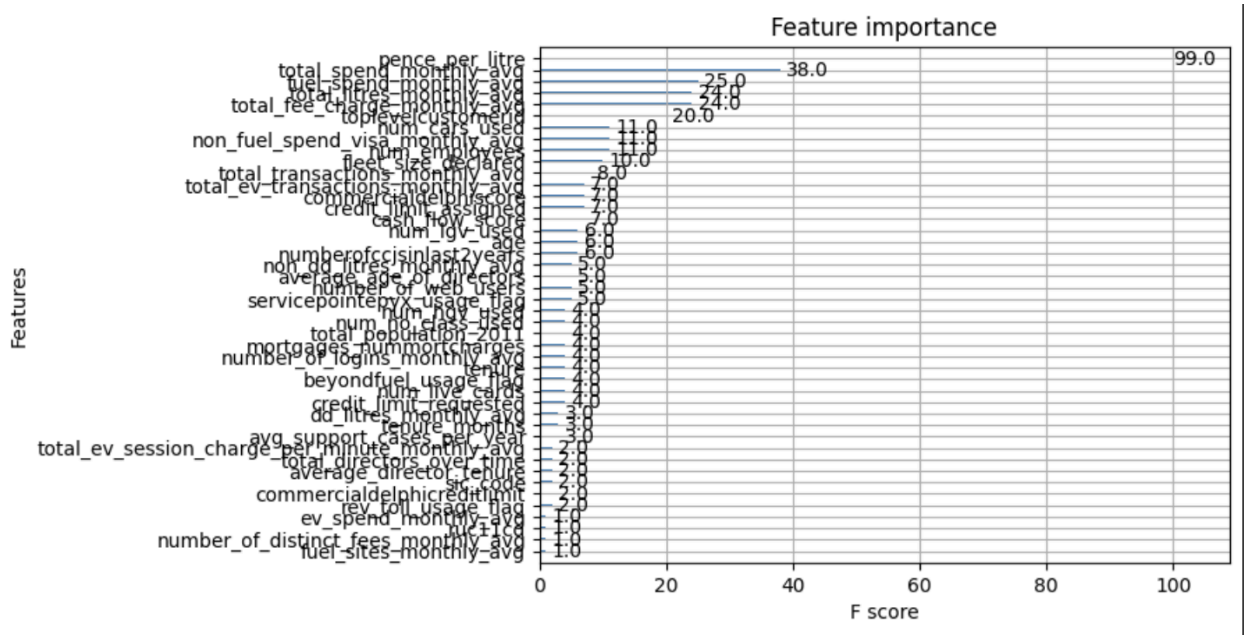
Analysis of the 'Age' feature revealed a skewed distribution, with a significant portion of users clustered below the age of 20.



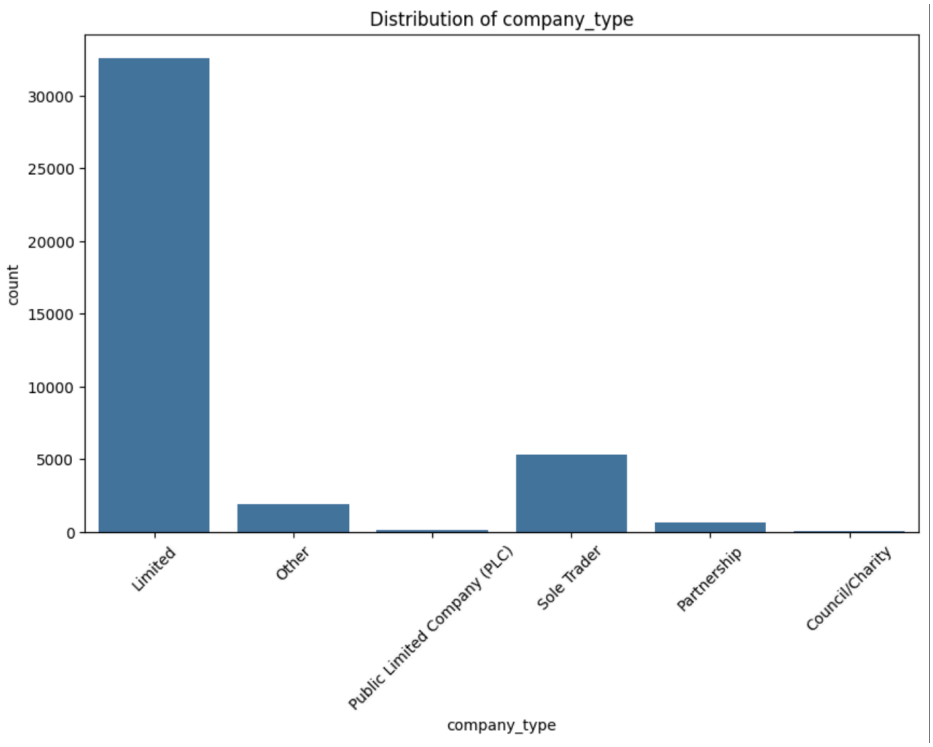
Correlation analysis identified 15 features with strong correlations (above 0.5) with 'total_rev_monthly_avg', providing valuable insights into potential predictors of revenue.

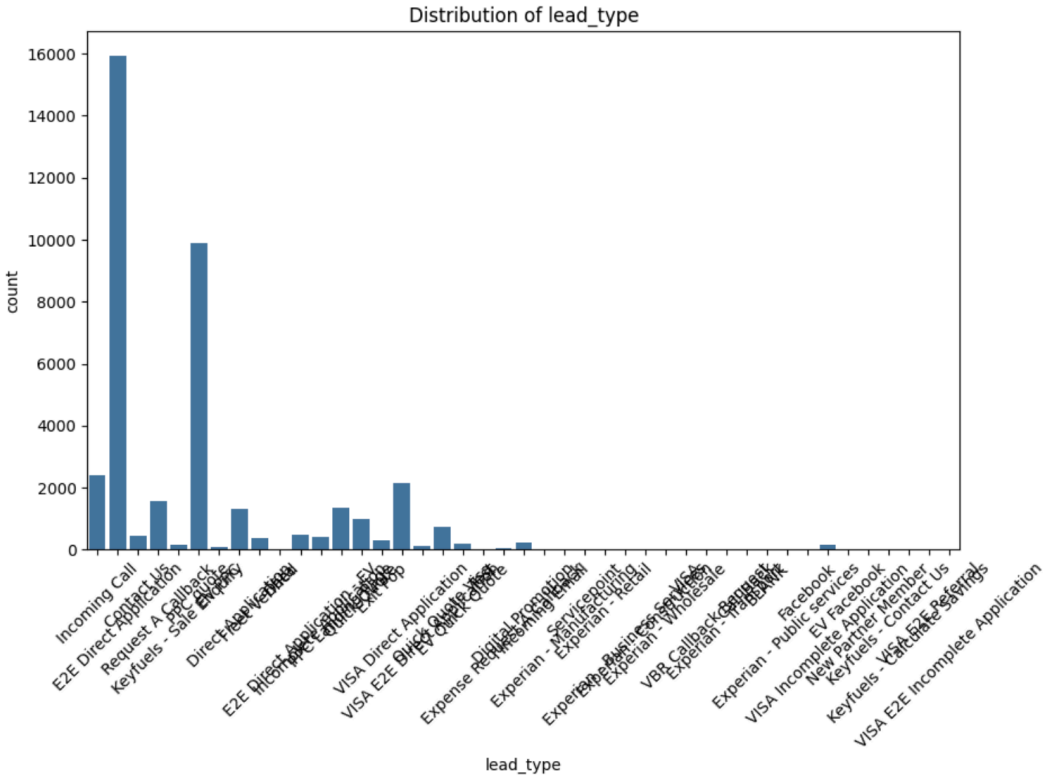


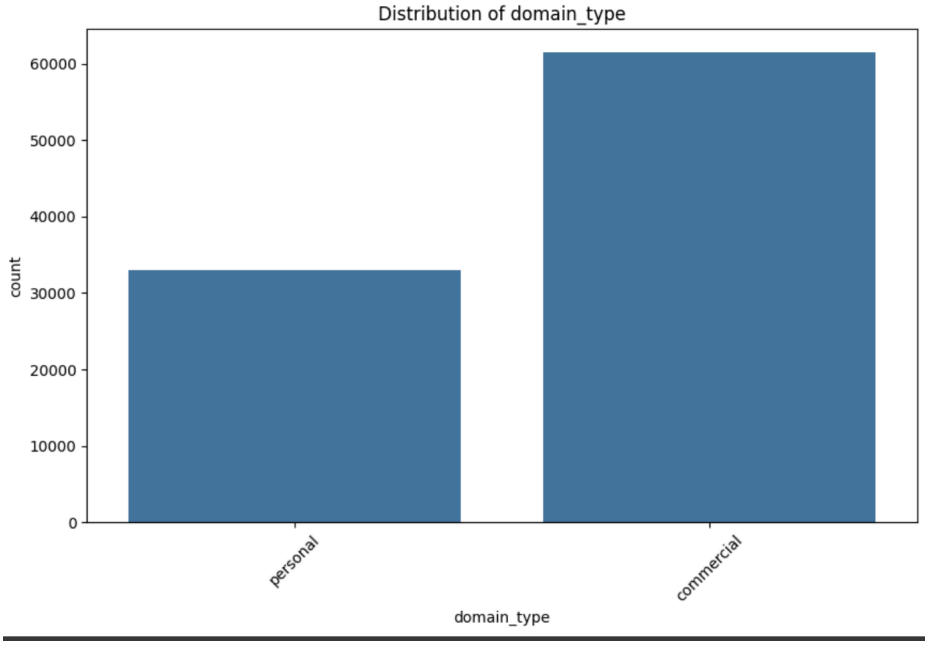
Additionally, feature importance evaluation using the XG Boost method yielded results consistent with correlation findings.



Moreover, distribution analysis based on 'company_type', 'lead_type', and 'domain_type' highlighted notable trends in company characteristics.




Distribution of lead_type
count
lead_type
Incoming Call
VISA Direct Application
VISA E2E Incomplete Application


Distribution of domain_type
count
domain_type
personal
commercial

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

```
data_mart_df_V3['is_holiday'].value_counts()
```

```
is_holiday
0    76455
1      169
Name: count, dtype: int64
```

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.

$$\text{Probability} = \sum_{Model\ 1}^{Model\ n} (Model\ Score * Model\ Accuracy)$$

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.