ACTL 3143 Assignment
Part 1 Report –Audrey Chang

Table of Contents

1. Problem Specification **2**

2. Data Collection **2**

3. Exploratory Data Analysis (EDA) **2**

3.1 Target Distribution and Class Imbalance 2
3.2 Missing Values and Anomalies 3
3.3 Correlation Analysis of Numeric Variables 3
3.4 Categorical Feature Patterns 3
3.5 Additional Observations (See Appendix) 4

4. Data Preprocessing **4**

4.1 Initial Data Cleaning: Target Integrity and Duplicates 4
4.2 Feature Engineering and Redundancy Removal 5
4.3 Handling Outliers and Missing Values 5
4.4 Encoding High-Cardinality Categorical Features: ServiceArea 6
4.5 Train/Validation/Test Split 6
4.6 Data Dictionary (Selected Features) 6
4.7 Preparation for Modeling 7

5. Baseline Model and Metrics **7**

5.1 Model Choice 7
5.2 Training Procedure 7
5.3 Validation Performance 7
5.4 Test Set Performance 8
5.5 Interpretation and Insights 8
5.6 Summary 8
3.5 Additional Observations 10

1. Problem Specification

In the telecommunications industry, customer churn—when subscribers discontinue their service—poses a persistent challenge to profitability and long-term sustainability. Retaining existing customers is typically more cost-effective than acquiring new ones, making churn prediction a critical strategic focus for telecom providers(Saleh, 2023). Increased market liberalization and number portability have further intensified competition, enabling customers to switch providers with minimal friction.
This project investigates the problem of predicting customer churn using supervised machine learning techniques. Specifically, I formulate churn prediction as a binary classification task: determining whether a customer will churn (1) or remain subscribed 0) in the subsequent billing cycle. Model performance is evaluated using AUC-ROC, F1 Score, recall, precision, and accuracy.
Model performance will be evaluated using AUC-ROC and F1 Score as primary metrics, given the moderate class imbalance (~28.8%). Precision, recall, and accuracy will also be reported for a more complete view of performance.

2. Data Collection

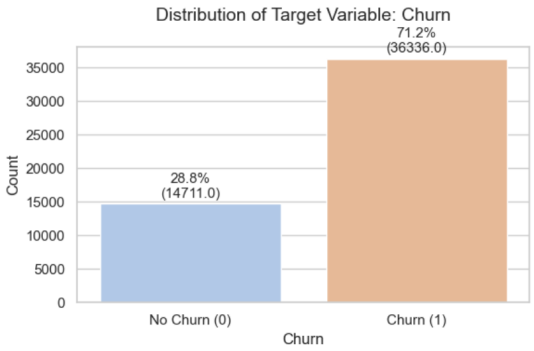
The dataset used in this project is the Cell2Cell: The Churn Dataset, originally published by the Teradata Center for Customer Relationship Management at Duke University and publicly available on Kaggle. This real-world dataset contains 71,047 customer records and 58 variables, capturing customer demographics, account tenure, service usage, billing information, and interaction history. Approximately 20,000 records belong to a designated holdout set without churn labels. These records were excluded from analysis to ensure supervised learning could be applied correctly. After filtering out unlabeled entries and removing duplicate rows, the working dataset comprises 51,047 observations and 52 cleaned and engineered features. A detailed data dictionary is provided in the Appendix.
This dataset is derived from actual operational data collected by the Teradata Center for Customer Relationship Management at Duke University in collaboration with a telecommunications provider. While all personally identifiable information has been removed, the variable definitions, distributions, and anomalies are consistent with real-world business data rather than synthetic or simulated records.

3. Exploratory Data Analysis (EDA)

This section summarizes key patterns, anomalies, and early insights observed in the raw dataset before preprocessing. The analysis informed the feature engineering and cleaning strategies adopted in subsequent steps.

3.1 Target Distribution and Class Imbalance

The target variable, Churn, indicates whether a customer discontinued service in the next billing cycle. Originally encoded as “Yes” and “No,” it was recoded to binary values (1 = churned, 0 = retained). Approximately **28.8% of customers** in the dataset had churned. This represents a **moderate class imbalance**, which can bias standard accuracy metrics toward the majority class. Accordingly, downstream model evaluation will emphasize AUC-ROC, F1 Score, and recall, as these better capture the trade-off between detecting churners and avoiding false positives.



3.2 Missing Values and Anomalies

Several features contain missing or invalid entries:(See Appendix)

- Household Age (AgeHH1, AgeHH2): ~900 missing values, likely due to non-responses.
- Usage Volatility Metrics (PercChangeMinutes, PercChangeRevenues): ~360 missing entries each, suggesting incomplete billing history.
- Billing Metrics (MonthlyRevenue, MonthlyMinutes, TotalRecurringCharge): ~150 missing entries each.

Additionally, some numerical columns contained implausible negative values (e.g., CurrentEquipmentDays as low as –5). These anomalies were flagged for imputation and correction during data preprocessing.

3.3 Correlation Analysis of Numeric Variables

To assess linear relationships between numeric features and churn, I computed Pearson correlation coefficients. While most variables exhibited only weak linear associations with churn, a few notable trends emerged:

- **Customer Service Interactions** (e.g., *RetentionCalls*) showed a positive correlation with churn, indicating that frequent support calls may reflect dissatisfaction or cancellation intent. These are valuable early signals for retention targeting.
- **Usage Patterns** like *MonthlyMinutes* and *TotalRecurringCharge* correlated negatively with churn, suggesting that high-engagement users are more loyal and reliant on the service—making them strong candidates for loyalty programs.
- **Equipment Tenure** (*CurrentEquipmentDays*) had one of the strongest positive correlations with churn. Long device usage may signal neglect or outdated technology, presenting an opportunity for proactive upgrade campaigns.

Although most individual features showed weak linear links to churn, combining behavioral, usage, and tenure variables is essential. For instance, long equipment tenure plus low usage may indicate high churn risk, while high usage may counteract that risk.Overall, this analysis reinforces two key insights:

- Churn prediction requires modeling feature interactions rather than relying on single variables.
- Segment-specific strategies such as upgrades for low-usage, long-tenure users and rewards for heavy users can help reduce churn.

These findings guided the feature engineering and modeling decisions in subsequent phases of the project.

3.4 Categorical Feature Patterns

Several key categorical and ordinal features revealed non-trivial associations with churn:

- **Credit Rating**: Churn does not increase monotonically with worsening credit. Customers in the top credit tiers (categories 1–3) displayed higher churn (~30%) compared to mid-tier segments

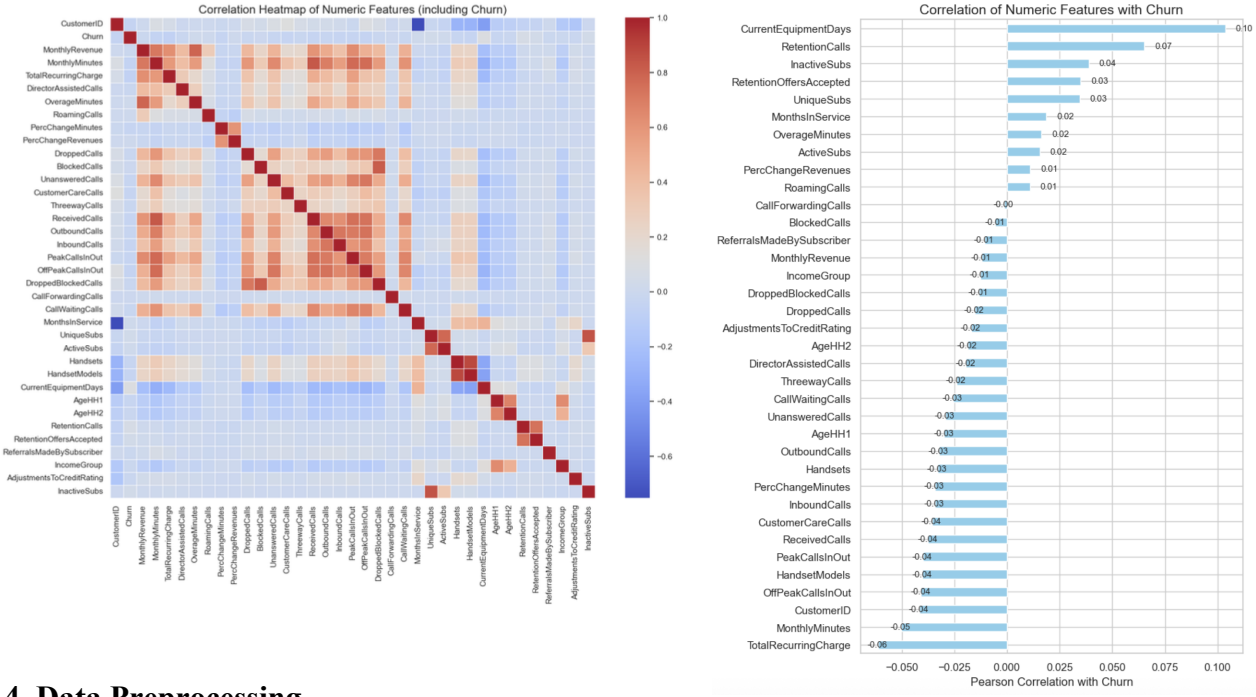
(~22%), suggesting that more financially reliable customers may also be more selective and prone to switching providers.

- **Equipment Tenure**: Binned analysis of CurrentEquipmentDays showed churn increasing beyond 300 days of device usage, indicating that lack of upgrades may drive dissatisfaction.
- **Months in Service**: A U-shaped churn pattern was observed. Very new customers (<10 months) churn less (~15%), but churn spikes (~42%) around the 10–12 month mark—likely reflecting contract renewals or the expiration of promotional periods.
- **Monthly Minutes**: Low-usage customers (< 90 minutes per month) had the highest churn (~35%), while higher-usage groups were progressively more loyal.
- **Overage Minutes**: Customers with higher overage minutes showed increasing churn, suggesting billing friction as a driver of attrition.
- **Handset Web Capability**: Customers with older, non-web-capable devices churned at ~37%, compared to ~28% for those with modern devices, highlighting technology adoption as an important factor.

These patterns underscore that churn is driven by a combination of usage intensity, tenure, billing dynamics, and device characteristics rather than any single factor in isolation.

3.5 Additional Observations (See Appendix)

- **Outliers**: Many numeric features exhibited long right tails, reinforcing the need for robust preprocessing techniques (e.g., median imputation, scaling).
- **High Cardinality**: Features such as ServiceArea have a large number of categories, motivating the use of smoothed target mean encoding to capture regional churn trends without inflating dimensionality.
- **Ambiguous Categories:** Variables like HandsetPrice contain substantial "Unknown" entries, which were earmarked for specialized encoding strategies.



4. Data Preprocessing

4.1 Initial Data Cleaning: Target Integrity and Duplicates

To prepare the dataset for modeling, I performed the following cleaning steps:

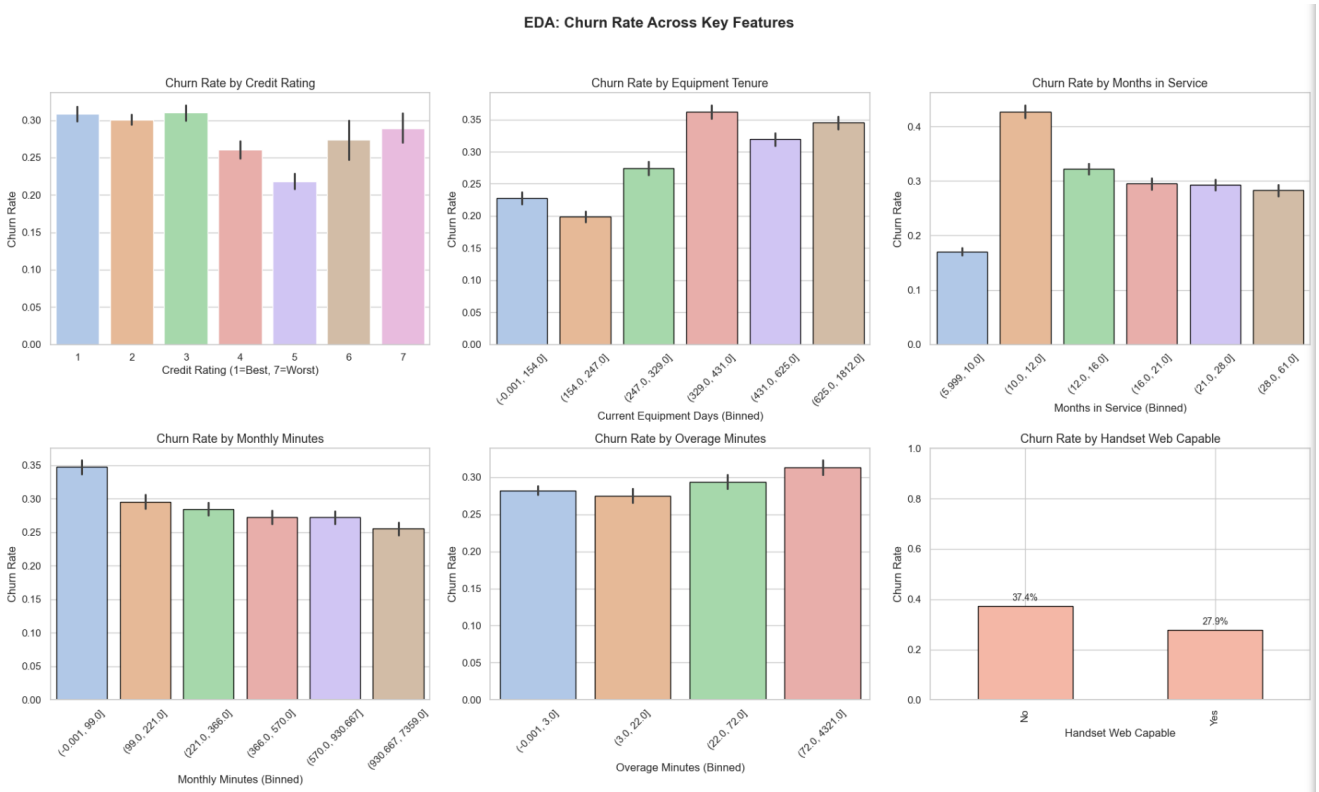
- Removal of Unlabeled Records: The raw dataset contained approximately 71,000 customer records, but around 20,000 of these belonged to a separate holdout set without churn labels. These observations were excluded to ensure supervised learning could be applied appropriately.
- Removed Duplicates: Duplicate rows were identified and deleted to avoid bias and data leakage.
- Negative and Invalid Values: Implausible negative entries were replaced with missing indicators (NaN) for later imputation rather than discarding the entire record.

4.2 Feature Engineering and Redundancy Removal

Several variables were dropped due to redundancy or low information content. I removed the customer identifier (CustomerID) since it carries no predictive value. Redundant or deterministic variables—such as NotNewCellphoneUser (which is complementary to NewCellphoneUser) and BlockedCalls (which overlaps with DroppedCalls) were excluded for parsimony. Columns like CallForwardingCalls, which had extremely low variance and near-zero correlation with churn, were also discarded.
To address multicollinearity, I examined highly correlated variables. For instance, UniqueSubs and ActiveSubs showed a Pearson correlation of 0.78. I retained the more interpretable engineered feature, InactiveSubs = UniqueSubs - ActiveSubs, and dropped the originals. Similarly, HandsetModels was strongly correlated with Handsets and was removed.

4.3 Handling Outliers and Missing Values

During exploratory analysis, I identified invalid values in several numeric columns. In particular, features such as MonthlyRevenue, TotalRecurringCharge, and CurrentEquipmentDays contained implausible negative values. Rather than deleting these records, I treated these values as missing (i.e., set them to NaN) and imputed them later using median values from the training set.



4.4 Encoding High-Cardinality Categorical Features: ServiceArea

A notable categorical feature, ServiceArea, had high cardinality. To encode it without introducing excessive dimensionality, I applied smoothed target mean encoding, replacing each region with a weighted average churn rate. This approach captures underlying geographic churn patterns while controlling for overfitting in areas with limited data. I used a smoothing parameter of 10 and filled unseen regions with the global mean of the training data set.

4.5 Train/Validation/Test Split

After all cleaning and feature engineering, I partitioned the data into training (60%), validation (20%), and test (20%) sets using a stratified split to preserve the original churn distribution (~28.8%). Crucially, the data was split prior to any imputation or scaling to prevent data leakage. This ensured that statistics computed during preprocessing (e.g., median imputation values) were based only on training data.

4.6 Data Dictionary (Selected Features)

Below is the three-column data dictionary for the key variables used in modeling. The full dictionary appears in Appendix A. These selected features represent the most important categories: customer usage, billing behavior, device characteristics, and service interactions. Additional derived and support features were used in training and are listed in the appendix.

Variable Name	Description	Datatype
Churn	Whether the customer has churned (1 = churned, 0 = retained)	Numeric (Binary)
CreditRating	Ordinal credit score category (1 = Highest, 4 = Medium). Indicates financial trustworthiness	Ordinal Categorical
IncomeGroup	Household income level from 1 (lowest) to 9 (highest)	Ordinal Categorical
CurrentEquipmentDays	Days the current device has been used	Numeric
MonthsInService	Total number of months the customer has been with the provider	Numeric
MonthlyRevenue	Monthly amount paid by the customer	Numeric
MonthlyMinutes	Total monthly voice minutes used	Numeric
OverageMinutes	Minutes exceeding the customer's plan limit	Numeric
TotalRecurringCharge	Monthly recurring charges excluding overages	Numeric
RetentionOffersAccepted	Number of accepted offers aiming to prevent churn	Numeric
HandsetWebCapable	Whether the handset supports web access	Nominal Categorical

HandsetPrice_Clean	Numeric handset price tier after parsing; “Unknown” handled separately.	Numeric
HandsetPrice_Unknown	Flag indicating whether handset price was originally “Unknown”.	Numeric
InactiveSubs	Engineered feature: Number of inactive subscriptions (UniqueSubs – ActiveSubs)	Numeric
ServiceArea	Customer’s geographical service region. This nominal categorical feature was encoded using smoothed target mean encoding to capture churn risk while reducing dimensionality.	Encoded Numeric (was Nominal)

4.7 Preparation for Modeling

To streamline training, I implemented two distinct preprocessing pipelines: one for the baseline model and one for deep learning. Both pipelines handled missing values with median imputation for numeric fields and mode imputation for categoricals. Categorical features were encoded using a combination of ordinal encoders and one-hot encoders, depending on their nature.
The deep learning pipeline included standard scaling to normalize feature distributions for neural network convergence. Due to solver convergence issues, the baseline pipeline was also updated to include standard scaling, improving numerical stability.
Both pipelines were fitted exclusively on the training set and serialized with joblib for reproducibility.

5. Baseline Model and Metrics

To establish a benchmark for comparison, I trained a logistic regression model as the baseline for churn prediction. This simple model offers interpretability and provides a reference point for evaluating more complex deep learning architectures.

5.1 Model Choice

Logistic regression was selected due to its simplicity, transparency, and effectiveness in binary classification problems. It also produces calibrated class probabilities, allowing for robust evaluation with metrics such as AUC-ROC and F1 score.
Given the class imbalance (28.8% churn), I applied class_weight='balanced' to ensure the model placed equal emphasis on both classes during training.

5.2 Training Procedure

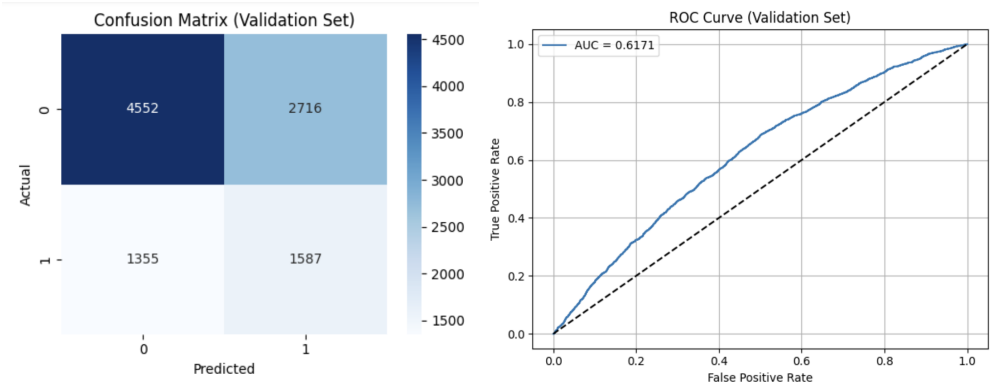
The logistic regression model was trained on the processed training set (X_train_base, y_train) and evaluated on a hold-out validation set (X_val_base, y_val). Key training parameters included:

- Solver: lbfgs (robust for medium-scale data)
- Max Iterations: 3000 (to ensure convergence)
- Class Weight: 'balanced'
- Random State: 42 (for reproducibility)

Standard scaling was applied to all numeric features to improve model convergence and performance.

5.3 Validation Performance

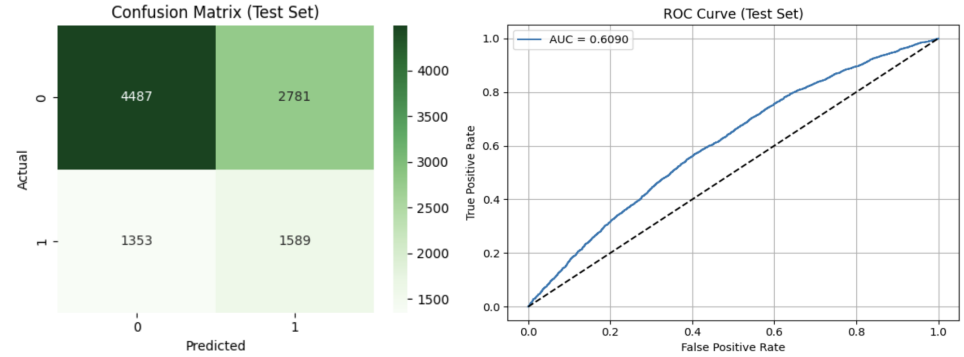
The model achieved moderate predictive performance on the validation set:



The confusion matrix shows that while non-churners are predicted more reliably, the model misses many actual churners (false negatives). The ROC curve indicates modest discriminatory power, and the precision-recall curve highlights the challenge of maintaining precision as recall increases.(**precision-recall curve** and **full classification report** in Appendix A.)

5.4 Test Set Performance

On the test set (X_test_base, y_test), model performance remained consistent:



These results confirm that the model generalizes reasonably well. The slight drop in AUC from validation (0.6171) to test (0.6090) indicates limited but stable learning. The consistency across sets suggests no overfitting, but also that the model's ability to detect churners remains constrained by its linear nature.

5.5 Interpretation and Insights

While logistic regression establishes a reproducible and interpretable baseline, its limited performance reinforces the complexity of the churn problem. The model is able to capture some churn-related signals but struggles to separate classes cleanly. This is expected due to overlapping feature distributions and nonlinear patterns observed during EDA. The model's modest recall for the positive class (~53–54%) highlights the need for improvement, particularly in minimizing false negatives. This supports the case for exploring more expressive deep learning models that can capture complex feature interactions.

5.6 Summary

The logistic regression baseline achieves an ROC AUC of approximately 0.61–0.62, with moderate accuracy (60%) and recall (53–54%). While simple and interpretable, it is not sufficient for high-precision churn prediction. These results provide a realistic benchmark against which to compare upcoming deep learning models.
The trained model was saved as logistic_baseline_model.pkl for reproducibility.

References

jpacse. (2020). *Datasets for churn telecom* [Data set]. Kaggle. https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom

S. Saleh, S. Saha, Customer retention and churn prediction in the telecommunication industry: a case study on a Danish University, Soc. Netw. Anal. Appl. Sci. 5 (2023) 173, https://link.springer.com/article/10.1007/s42452-023-05389-6

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). *Explaining customer churn prediction in telecom industry using tabular machine learning models*. Machine Learning with Applications, 17, 100567. https://doi.org/10.1016/j.mlwa.2024.100567

Beeharry, Y., & Tsokizep Fokone, R. (2022). *Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry*. Concurrency and Computation: Practice and Experience, 34(4), e6627. https://doi.org/10.1002/cpe.6627

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). *Customer churn prediction in telecom sector using machine learning techniques*. Results in Control and Optimization, 14, 100342. https://doi.org/10.1016/j.rico.2023.100342

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing, 104*(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

Asif, D., Arif, M. S., & Mukheimer, A. (2025). A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results in Engineering, 26*, 104629. https://doi.org/10.1016/j.rineng.2025.104629

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Appendix A3. EDA

3.2 Missing Values and Anomalies

- **missing entries**

```
Missing values in columns:
AgeHH1                   909
AgeHH2                   909
PercChangeMinutes        367
PercChangeRevenues       367
MonthlyRevenue           156
MonthlyMinutes           156
TotalRecurringCharge     156
DirectorAssistedCalls    156
OverageMinutes           156
RoamingCalls             156
ServiceArea               24
Handsets                   1
HandsetModels              1
CurrentEquipmentDays       1
dtype: int64
```

3.5 Additional Observations

- **Ambiguous Categories:**

```
HandsetPrice category distribution:
HandsetPrice
Unknown    28914
30          7268
150         4086
130         2090
80          1938
10          1916
60          1761
200         1257
100         1226
40           247
400           46
250           19
300           13
180           10
500            8
240            6
Name: count, dtype: int64
```

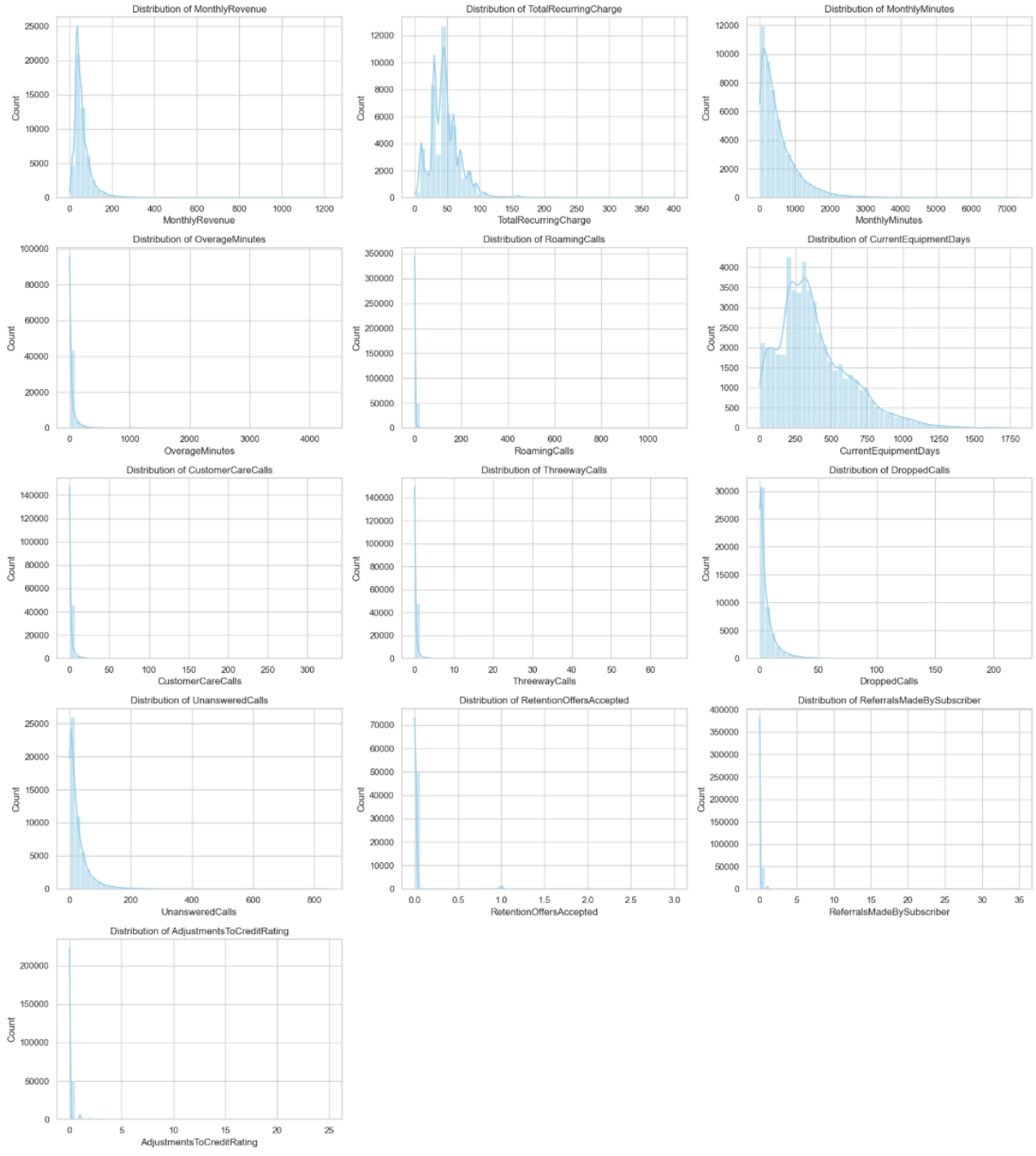
- **implausible negative values**

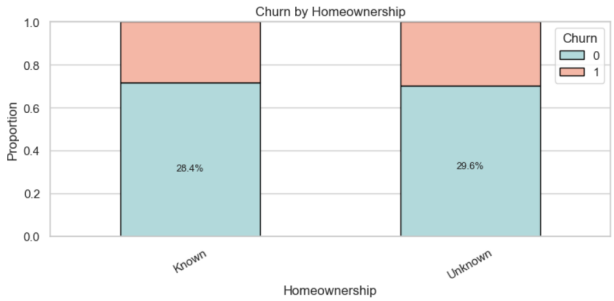
```
MonthlyRevenue has 3 negative values
TotalRecurringCharge has 8 negative values
CurrentEquipmentDays has 76 negative values
```

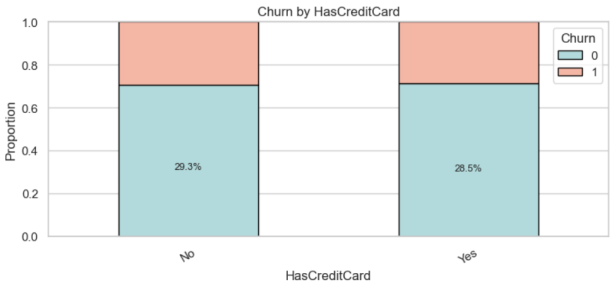
- High Cardinality:

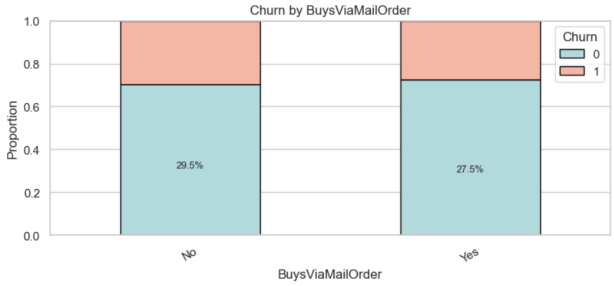
```
ServiceArea
NYCBRO917    1679
HOUHOU281    1508
DALDAL214    1496
NYCMAN917    1177
APCFCH703     781
DALFTW817     780
SANSAN210     721
APCSIL301     666
SANAUS512     610
SFROAK510     605
Name: count, dtype: int64
    ServiceArea  Count
0     NYCBRO917   1679
1     HOUHOU281   1508
2     DALDAL214   1496
3     NYCMAN917   1177
4     APCFCH703    781
..          ...    ...
743   OHIRAV330      1
744   KCYHES316      1
745   AIRGEO843      1
746   AIRNEW803      1
747   NCRDNN910      1

[748 rows x 2 columns]
```

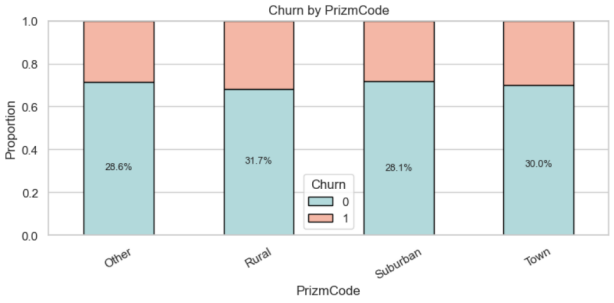
- **Outliers**:

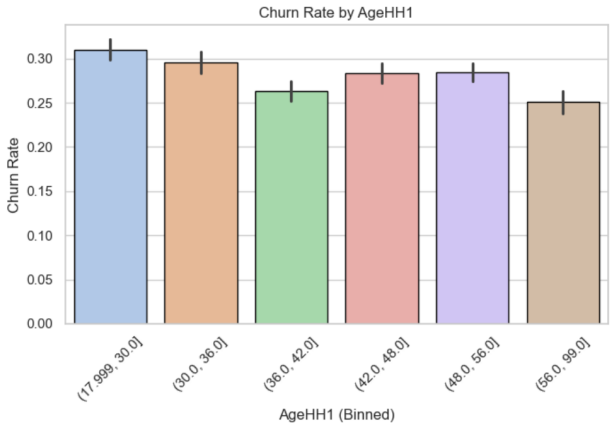

Distribution of MonthlyRevenue
Count
MonthlyRevenue
Distribution of TotalRecurringCharge
Count
TotalRecurringCharge
Distribution of MonthlyMinutes
Count
MonthlyMinutes
Distribution of OverageMinutes
Count
OverageMinutes
Distribution of RoamingCalls
Count
RoamingCalls
Distribution of CurrentEquipmentDays
Count
CurrentEquipmentDays
Distribution of CustomerCareCalls
Count
CustomerCareCalls
Distribution of ThreewayCalls
Count
ThreewayCalls
Distribution of DroppedCalls
Count
DroppedCalls
Distribution of UnansweredCalls
Count
UnansweredCalls
Distribution of RetentionOffersAccepted
Count
RetentionOffersAccepted
Distribution of ReferralsMadeBySubscriber
Count
ReferralsMadeBySubscriber
Distribution of AdjustmentsToCreditRating
Count
AdjustmentsToCreditRating

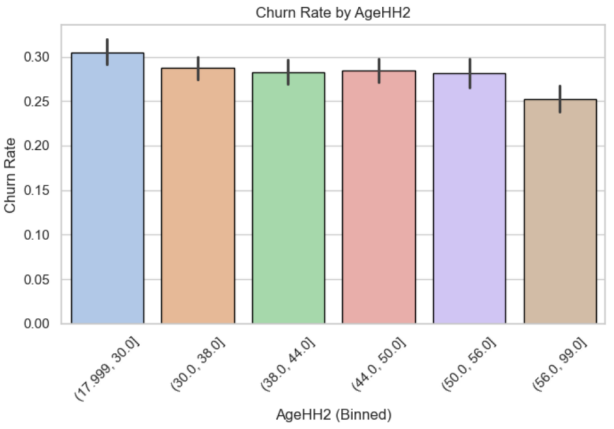

Churn by Homeownership
Proportion
Churn
0
1
28.4%
29.6%
Known
Unknown
Homeownership

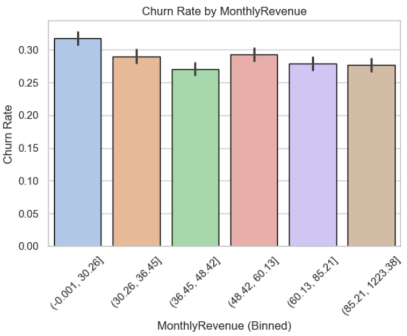

Churn by HasCreditCard
Proportion
Churn
0
1
29.3%
28.5%
No
Yes
HasCreditCard

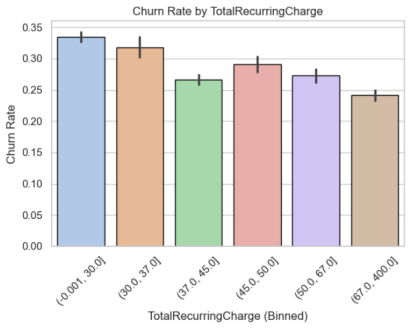

Churn by BuysViaMailOrder
Proportion
Churn
0
1
29.5%
27.5%
No
Yes
BuysViaMailOrder

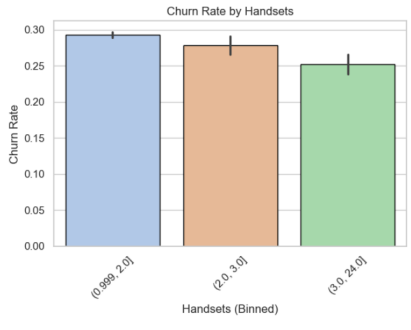

Churn by PrizmCode
Proportion
28.6%
31.7%
28.1%
30.0%
Churn
0
1
Other
Rural
Suburban
Town
PrizmCode

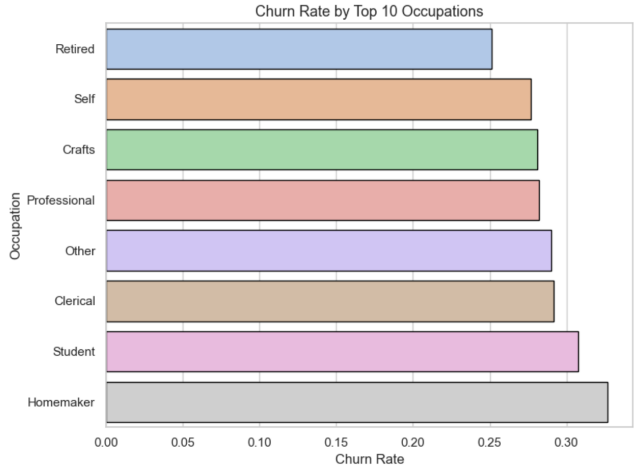

Churn Rate by AgeHH1
Churn Rate
(17.999, 30.0]
(30.0, 36.0]
(36.0, 42.0]
(42.0, 48.0]
(48.0, 56.0]
(56.0, 99.0]
AgeHH1 (Binned)

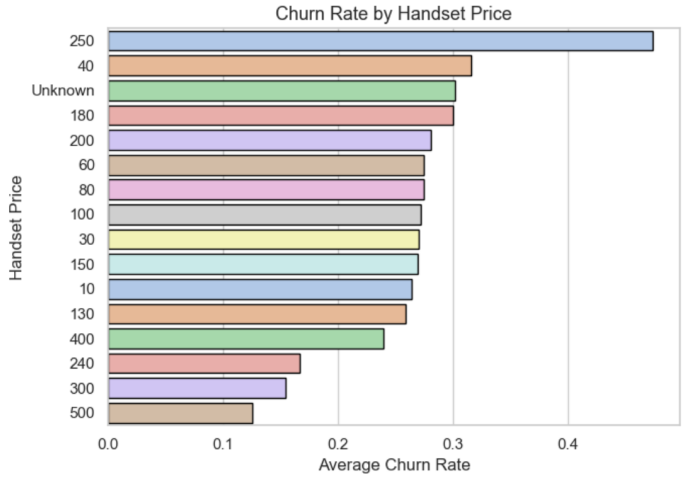

Churn Rate by AgeHH2
Churn Rate
(17.999, 30.0]
(30.0, 38.0]
(38.0, 44.0]
(44.0, 50.0]
(50.0, 56.0]
(56.0, 99.0]
AgeHH2 (Binned)


Churn Rate by MonthlyRevenue
Churn Rate
(-0.001, 30.26]
(30.26, 36.45]
(36.45, 48.42]
(48.42, 60.13]
(60.13, 85.21]
(85.21, 1223.38]
MonthlyRevenue (Binned)


Churn Rate by TotalRecurringCharge
Churn Rate
(-0.001, 30.0]
(30.0, 37.0]
(37.0, 45.0]
(45.0, 50.0]
(50.0, 67.0]
(67.0, 400.0]
TotalRecurringCharge (Binned)


Churn Rate by Handsets
Churn Rate
(0.999, 2.0]
(2.0, 3.0]
(3.0, 24.0]
Handsets (Binned)





Data dictionary

Variable Name	Description	Datatype
Churn	Whether the customer has churned (Yes = churned, No = retained)	Numeric (Binary)
InactiveSubs	Engineered feature: Number of inactive subscriptions (UniqueSubs – ActiveSubs)	Numeric
ServiceArea	Customer's geographical service region. This nominal categorical feature was encoded using smoothed target mean encoding (smoothing parameter = 10), replacing each region with its weighted average churn rate. This approach preserves geographic churn patterns while avoiding high dimensionality from one-hot encoding.	Encoded Numeric (Originally Nominal Categorical)
Handsets	Total number of handsets on account	Numeric
HandsetPrice_Clean	Numeric handset price tier after parsing; "Unknown" handled separately.	Numeric
HandsetPrice_Unknown	Flag indicating whether handset price was originally "Unknown".	Numeric

CurrentEquipmentDays	Days current device has been used	Numeric
HandsetRefurbished	Whether the customer's handset is a refurbished model (Yes = refurbished, No = new)	Nominal Categorical
HandsetWebCapable	Whether the handset supports web access	Nominal Categorical
ChildrenInHH	Whether there are children in the household	Nominal Categorical
AgeHH1	Age of primary household member	Numeric
AgeHH2	Age of secondary household member (if applicable)	Numeric
TruckOwner	Whether customer owns a truck	Nominal Categorical
RVOwner	Whether customer owns a recreational vehicle	Nominal Categorical
Homeownership	Whether customer owns their home	Nominal Categorical
BuysViaMailOrder	Whether customer shops via mail order	Nominal Categorical
RespondsToMailOffers	Whether customer responds to promotional mailings	Nominal Categorical
OptOutMailings	Whether customer opted out of promotional mailings	Nominal Categorical
NonUSTravel	Whether customer has traveled internationally	Nominal Categorical
OwnsComputer	Whether customer owns a personal computer	Nominal Categorical
HasCreditCard	Whether customer owns a credit card	Nominal Categorical
RetentionOffersAccepted	Number of retention offers accepted	Numeric
NewCellphoneUser	Whether customer is a new cellphone user	Nominal Categorical
ReferralsMadeBySubscriber	Referrals made by the customer	Numeric
IncomeGroup	Encoded household income bracket (1 = lowest income, 9 = highest income). This is an ordinal variable indicating relative income levels.	Ordinal Categorical
CreditRating	Ordinal credit score category where 1 = Highest, 2 = High, 3 = Good, and 4 = Medium. Reflects customer’s financial trustworthiness.	Ordinal Categorical
HandsetPrice	Price tier of the current handset	Ordinal Categorical
OwnsMotorcycle	Whether customer owns a motorcycle	Nominal Categorical
AdjustmentsToCreditRating	Number of changes to credit rating	Numeric
MadeCallToRetentionTeam	Whether customer made a call to retention	Nominal Categorical
PrizmCode	Customer's residential environment type: Suburban, Town, Rural, or Other. This feature reflects simplified socio-geographic segmentation.	Nominal Categorical
Occupation	Customer occupation category	Nominal Categorical
MaritalStatus	Indicates whether the customer is married (Yes), not married (No), or unknown. Simplified binary marital status with missing values labelled as Unknown.	Nominal Categorical

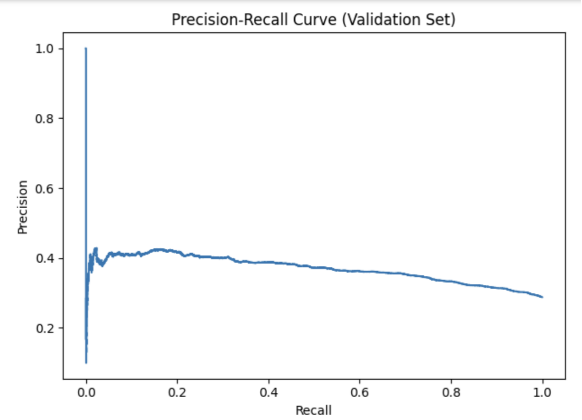
MonthlyRevenue	Monthly amount paid by the customer	Numeric
MonthlyMinutes	Total number of voice minutes used in the month	Numeric
TotalRecurringCharge	Monthly recurring charges excluding overages	Numeric
DirectorAssistedCalls	Number of calls assisted by a directory operator	Numeric
OverageMinutes	Minutes used beyond the customer's plan limit	Numeric
RoamingCalls	Number of calls made while roaming	Numeric
PercChangeMinutes	Percentage change in minutes used compared to previous month	Numeric
PercChangeRevenues	Percentage change in monthly revenue compared to previous month	Numeric
DroppedCalls	Number of dropped calls	Numeric
UnansweredCalls	Number of calls not answered	Numeric
CustomerCareCalls	Number of calls made to customer care	Numeric
ThreewayCalls	Number of three-way calls made	Numeric
ReceivedCalls	Number of calls received by the customer	Numeric
OutboundCalls	Number of calls made by the customer	Numeric
InboundCalls	Number of calls received from others	Numeric
PeakCallsInOut	Number of calls made or received during peak hours	Numeric
OffPeakCallsInOut	Number of calls made or received during off-peak hours	Numeric
CallWaitingCalls	Number of call waiting events	Numeric
MonthsInService	Number of months customer has been with the provider	Numeric

5. Baseline Model and Metrics

Classification Report (Validation Set):

```
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.77      0.63      0.69      7268
           1       0.37      0.54      0.44      2942

    accuracy                           0.60     10210
   macro avg       0.57      0.58      0.56     10210
weighted avg       0.65      0.60      0.62     10210

ROC AUC Score (Validation Set): 0.6171
```

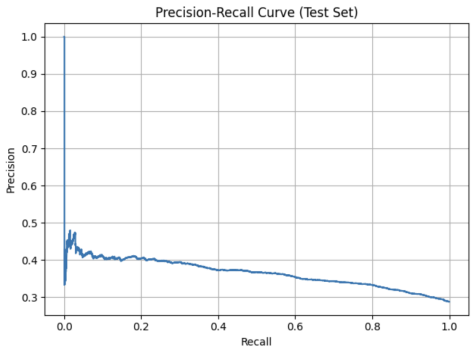
The precision-recall curve (**Validation Set**)



Classification Report (Test Set):

```
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.77      0.62      0.68      7268
           1       0.36      0.54      0.43      2942

    accuracy                           0.60     10210
   macro avg       0.57      0.58      0.56     10210
weighted avg       0.65      0.60      0.61     10210

ROC AUC Score (Test Set): 0.6090
```

The precision-recall curve (Test Set)



Appendix B – Generative AI Usage

OpenAI.(2025). *ChatGPT* (Jul 7 version) [Large language model]. https://chat.openai.com/chat

Generative AI tools were used throughout this assignment to support code development, debugging, and documentation. Specifically, ChatGPT (version 4) was used as a productivity assistant to accelerate coding tasks and clarify technical concepts. AI-generated outputs were not used verbatim in the report and were always reviewed, tested, and edited to suit the specific context of my dataset and modeling objectives.

Some examples of how Generative AI contributed to this project include:

- **Data Preprocessing Advice:** I asked ChatGPT how to handle features with high cardinality. In particular, I sought guidance on encoding techniques suitable for geographic categorical features like `ServiceArea`, where one-hot encoding would be impractical. Based on this, I applied *smoothed target mean encoding* to capture regional churn risk without inflating dimensionality.
- **Code Snippet Generation:** AI was used to generate Python code templates for tasks such as plotting binned churn rates, handling missing values, and provided guidance on structuring preprocessing pipelines.
- **Debugging and Refactoring:** I used ChatGPT to troubleshoot common errors in scikit-learn pipelines and confirm best practices for avoiding data leakage (e.g., ensuring data is split before imputation and scaling).
- **Writing and Editing Support:** AI provided suggestions for structuring sections of the report, such as framing the business problem.

Examples of prompts include:

- "What is the best way to encode a high-cardinality categorical variable without one-hot encoding all levels?"
- “What is a good way to stratify a train-test split when the target is imbalanced?”
- “How can I efficiently apply median imputation only on training data to avoid data leakage?”
- “Help me write a bar chart comparing churn rates across binned numeric features like MonthlyMinutes.”
- "Can you help me rephrase this section to be more concise and academic?"
- “Why am I getting ConvergenceWarning when using logistic regression with scikit-learn?”

All AI-assisted content was independently verified, rewritten where necessary, and adapted to ensure originality, accuracy. Overall, Generative AI was used responsibly to streamline development and enhance clarity, without replacing critical thinking, analytical reasoning, or domain understanding.