Automated EDA Report

Generated on: 2025-06-27 09:33:15

Table of Contents

Automated EDA Report

1. Executive Summary

This report presents a comprehensive Exploratory Data Analysis (EDA) of the dataset.

The original dataset contained 7043 rows and 21 columns.
After cleaning, the dataset had 7043 rows and 21 columns.

Feature engineering created 12 new features,
resulting in a final dataset with 7043 rows and 33 columns.

The problem was identified as a Binary Classification problem.

After evaluating 5 different models,
the best performing model was Pipeline(steps=[('prep',
ColumnTransformer(transformers=[('num',
Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')),
('scaler',
StandardScaler())]),
['SeniorCitizen', 'tenure',
'MonthlyCharges',
'TotalCharges',
'lagged_churn_rate_t1',
'rolling_avg_monthly_charges_3m',
'contract_paperless_billing_interaction',
'log_total_charges',
'monthly_charges_stan...
'OnlineSecurity',
'OnlineBackup',
'DeviceProtection',
'TechSupport', 'StreamingTV',
'StreamingMovies',
'Contract',
'PaperlessBilling',
'PaymentMethod',
'tenure_month_bin',
'contract_Month-to-month',
'contract_One year',
'contract_Two year'])])),
('model',
LogisticRegression(l1_ratio=0.5, max_iter=5000, n_jobs=-1,
penalty='elasticnet', random_state=42,
solver='saga'))]).

Automated EDA Report

2. Data Cleaning

The data cleaning process transformed the dataset from 7043 rows and 21 columns
to 7043 rows and 21 columns.

The following cleaning steps were performed:

Recommended Data Cleaning Steps:

Here are the recommended steps to clean and preprocess the provided dataset, tailored to the characteristics of the data and the user instructions:

1. **Convert Data Types**:

- Convert the `TotalCharges` column from `object` to `float64` since it contains numeric values. This is necessary for any numerical analysis or calculations.

```python
# Convert TotalCharges to float
df['TotalCharges'] = df['TotalCharges'].astype(float)
```

2. **Check for Missing Values**:

- Since there are no missing values in any column (0.00% missing), this step can be skipped. However, it's important to confirm this before proceeding.

3. **Impute Missing Values**:

- As per user instructions, handle missing values but since there are none, this step is not required.

4. **Remove Duplicate Rows**:

- Check for and remove any duplicate rows in the dataset to ensure data integrity.

```python
# Remove duplicate rows
df.drop_duplicates(inplace=True)
```

5. **Check for Outliers**:

- Although the user requested not to remove outliers, it is still beneficial to identify them for further analysis. This can be done using the interquartile range (IQR) method.

```python
# Identify outliers using IQR
```

```
Q1 = df['MonthlyCharges'].quantile(0.25)
Q3 = df['MonthlyCharges'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['MonthlyCharges'] < (Q1 - 3 * IQR)) | (df['MonthlyCharges'] > (Q3 + 3 * IQR))]
```

6. **Check for Unit Inconsistencies**:
   - Analyze numeric columns to ensure there are no unit inconsistencies (e.g., `TotalCharges` should be in the same unit as `MonthlyCharges`). Since `TotalCharges` is now converted to float, ensure that it aligns with the expected unit.
   - If any inconsistencies are found, standardize them accordingly.

7. **Analyze Data for Additional Cleaning Needs**:
   - Perform a general analysis of the data to determine if any additional cleaning steps are necessary. This includes checking for any unexpected values or inconsistencies in categorical columns.

8. **Final Review**:
   - Conduct a final review of the dataset to ensure all steps have been executed correctly and that the data is ready for analysis.

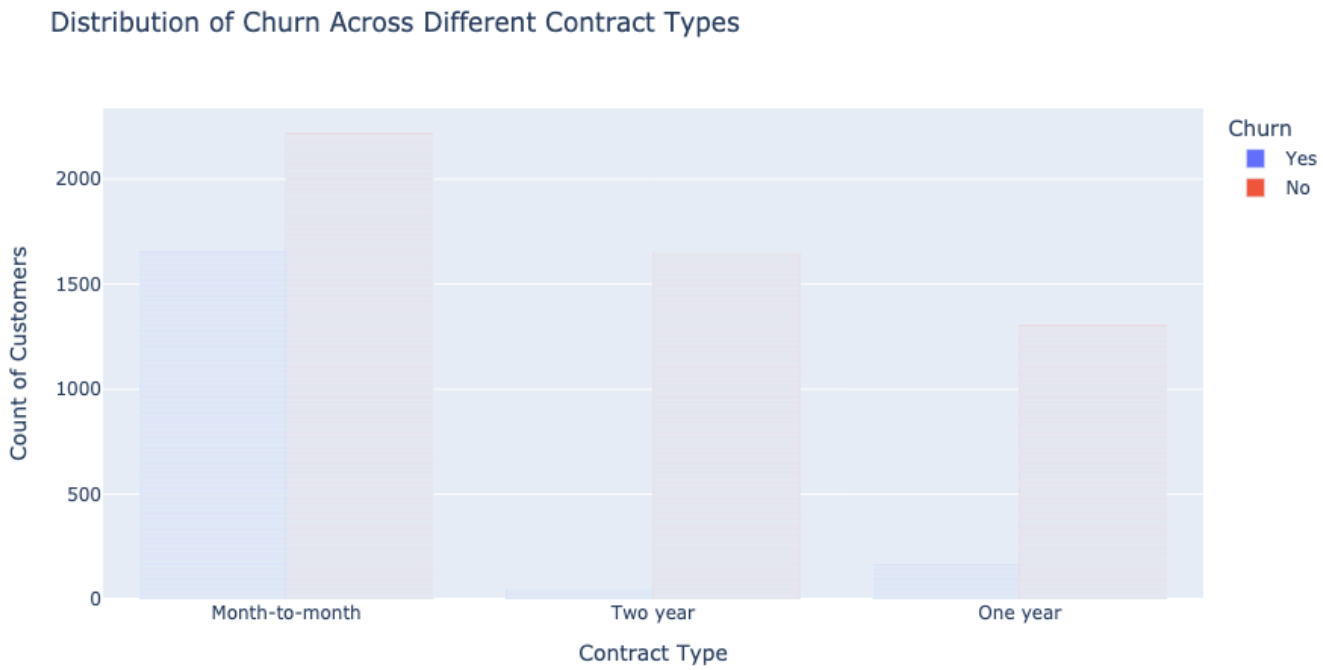
By following these steps, the dataset will be cleaned and preprocessed effectively, ensuring it is in a suitable format for further analysis.

Automated EDA Report

3. Data Visualization

The following visualizations provide key insights into the dataset patterns, distributions, and relationships.
Each visualization is accompanied by key observations that highlight important findings.

Churn Rate by Contract Type



Key Observations:

1. High Churn Rate in Month-to-Month Contracts

The data shows a significantly higher count of churn among customers with month-to-month contracts compared to those with one-year or two-year contracts. This suggests that customers on month-to-month contracts are more likely to leave, possibly due to the flexibility of not being tied to a long-term commitment. Businesses could consider offering incentives or discounts to month-to-month customers to encourage longer-term commitments and reduce churn.

2. Lower Churn in Long-term Contracts

Customers with one-year and two-year contracts exhibit a lower churn rate. This indicates that long-term contracts may provide a sense of stability or satisfaction, reducing the likelihood of customers leaving. Companies might focus on promoting the benefits of long-term contracts to new customers as a strategy to enhance customer retention.

3. Predominance of Month-to-Month Contracts

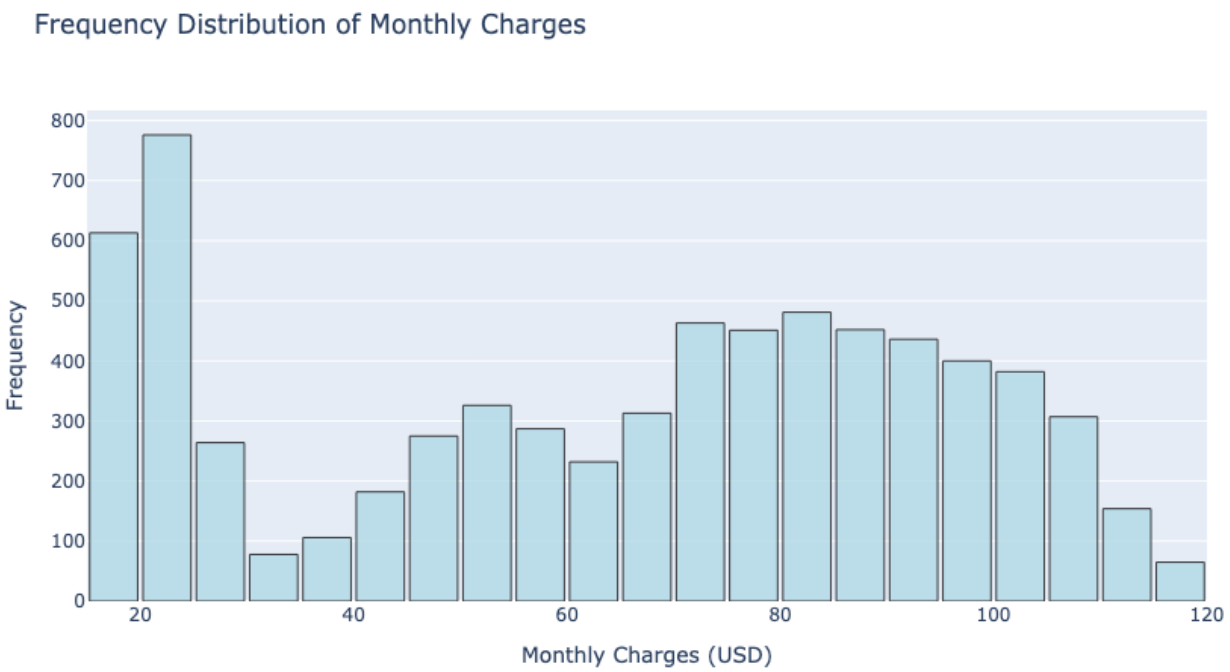
The majority of the data points are associated with month-to-month contracts, indicating that this is the most common contract type among customers. This trend highlights the preference for flexibility among customers. Businesses could explore enhancing the value proposition of month-to-month plans to cater to this preference while also considering ways to transition these customers to longer-term plans.

4. Potential for Targeted Retention Strategies

Given the high churn rate in month-to-month contracts, there is an opportunity to implement targeted retention strategies for these customers. Personalized offers, improved customer service, or loyalty programs could be effective in reducing churn rates. Understanding the specific reasons for churn in this group could further refine these strategies.

Automated EDA Report

Distribution of Monthly Charges



Key Observations:

1. Bimodal Distribution of Monthly Charges

The histogram of Monthly Charges shows a bimodal distribution with two significant peaks. The first peak occurs around $20-$30, while the second peak is observed around $70-$80. This suggests that there are two distinct groups of customers with different pricing plans or service levels. Understanding the characteristics of these groups could help in tailoring marketing strategies or service offerings.

2. Presence of High Monthly Charges Outliers

There are outliers in the Monthly Charges data, with some customers paying over $100 per month. These outliers could represent customers with premium service packages or additional features. Identifying these customers and understanding their needs could provide opportunities for upselling or cross-selling additional services.

3. Low Monthly Charges Cluster

A significant cluster of customers is paying low monthly charges, specifically in the range of $20-$30. This could indicate a basic service plan or a promotional offer. Analyzing this group could reveal insights into customer retention strategies and the effectiveness of entry-level pricing models.

4. Potential for Pricing Strategy Optimization

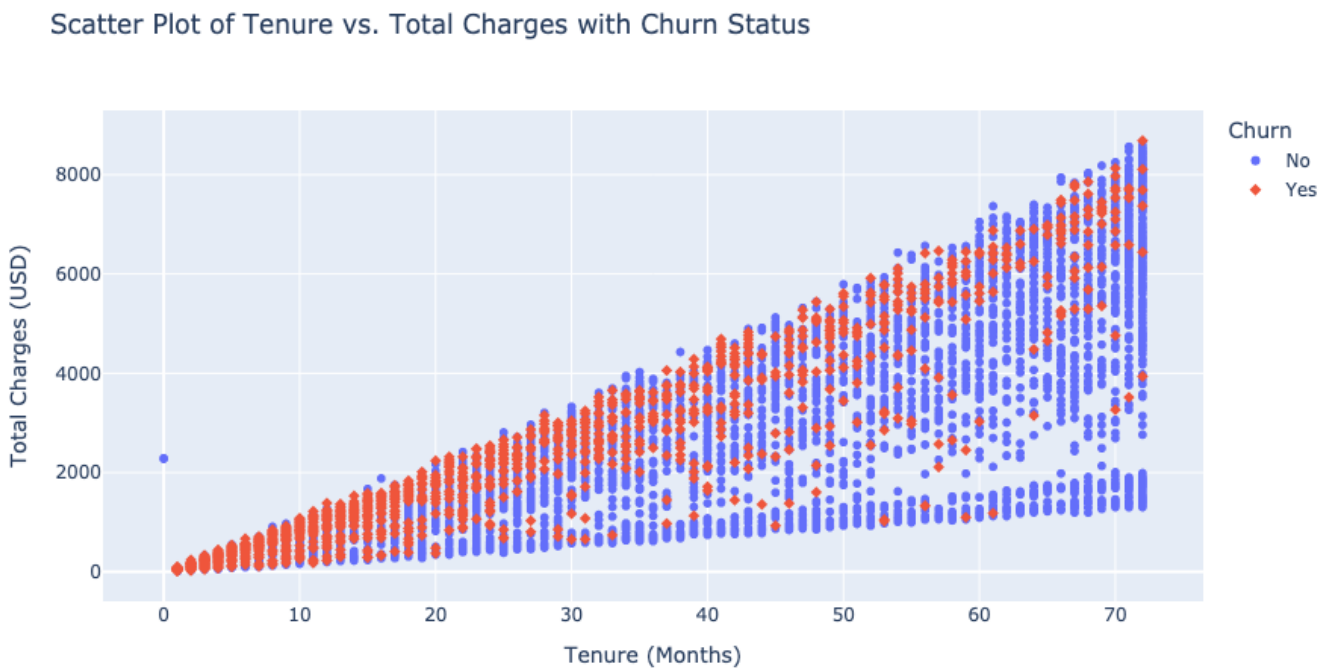
The distinct peaks in the distribution suggest that there might be room for optimizing pricing strategies. By analyzing the features and services associated with each peak, the company could adjust pricing tiers to better match customer preferences and maximize revenue.

5. Implications for Customer Segmentation

The distribution of Monthly Charges indicates clear segmentation opportunities. By categorizing customers based on their monthly charges, the company can develop targeted marketing campaigns and personalized service offerings to enhance customer satisfaction and loyalty.

Automated EDA Report

Tenure vs. Total Charges by Churn Status



Key Observations:

1. Higher Tenure Correlates with Lower Churn

The scatter plot reveals that customers with longer tenures tend to have lower churn rates. Most of the 'No' churn data points are clustered towards higher tenure values, indicating that customers who have been with the company longer are less likely to leave. This suggests that customer loyalty increases with tenure, and efforts to improve customer retention could focus on engaging newer customers to extend their tenure.

2. High Total Charges Among Churned Customers

The plot shows that customers who have churned ('Yes' churn status) often have higher total charges compared to those who have not churned. This is evident from the presence of 'Yes' churn data points at higher Total Charges values. This pattern suggests that high cumulative costs might be a factor contributing to customer churn, highlighting the need for pricing strategies or loyalty programs to mitigate this risk.

3. Low Tenure and Low Total Charges Among Churned Customers

There is a noticeable cluster of churned customers with both low tenure and low total charges. This indicates that some customers decide to leave shortly after beginning their service, possibly due to dissatisfaction or unmet expectations. Addressing initial customer experiences and satisfaction could help reduce churn in this segment.

4. Distinct Clusters of Churn Status

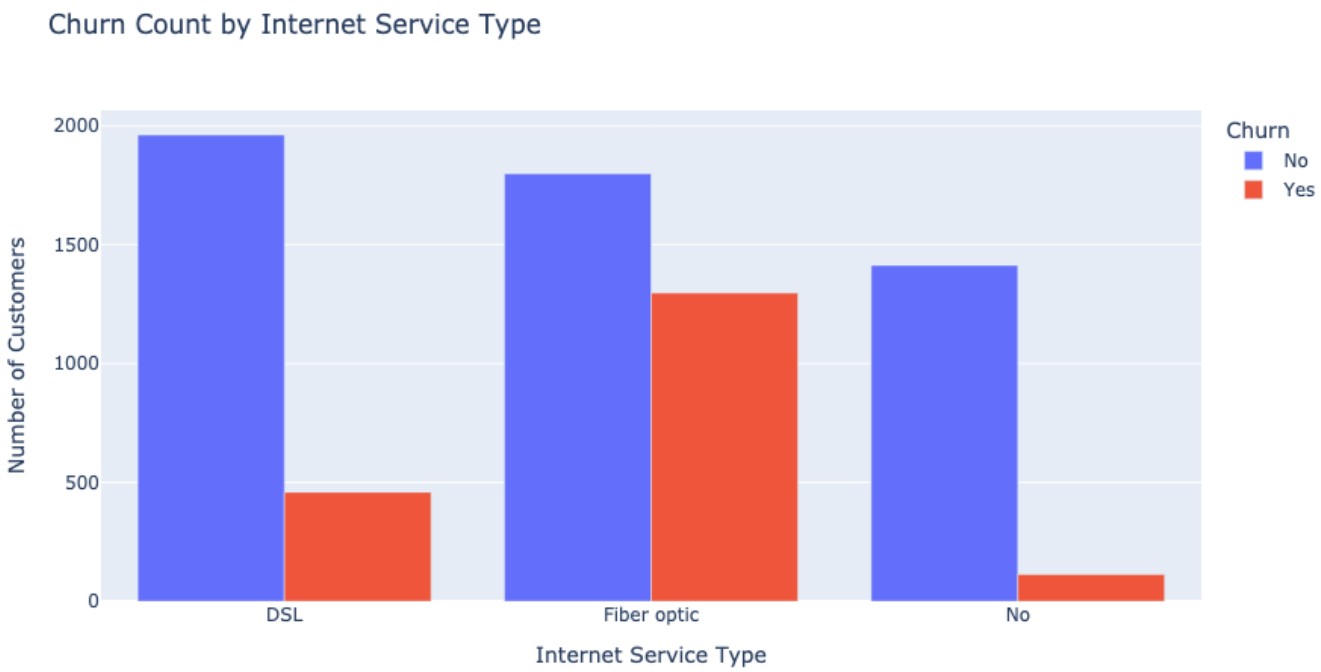
The scatter plot reveals distinct clusters based on churn status. 'No' churn customers are more spread out across a range of tenures and total charges, whereas 'Yes' churn customers are more concentrated in specific areas, particularly at lower tenures and higher total charges. This clustering suggests different customer profiles and behaviors, which could inform targeted retention strategies.

5. Potential Anomalies in Data

There are a few data points with very high total charges and low tenure that do not follow the general trend. These could be anomalies or outliers, possibly due to data entry errors or unique customer circumstances. Investigating these anomalies could provide insights into unusual customer behaviors or data quality issues.

Automated EDA Report

Churn Rate by Internet Service Type



Key Observations:

1. Higher Churn Rate for Fiber Optic Users

The bar chart indicates that customers with Fiber Optic internet service have a significantly higher churn rate compared to those with DSL or no internet service. The number of customers who churned ('Yes') is visibly larger for Fiber Optic users, suggesting potential dissatisfaction or issues with this service type. This insight could prompt the business to investigate the reasons behind the higher churn rate and address any service-related issues to improve customer retention.

2. DSL Users Show Lower Churn Rate

DSL internet service users exhibit a lower churn rate compared to Fiber Optic users. The count of customers who did not churn ('No') is higher for DSL users, indicating better customer retention. This could imply that DSL service meets customer expectations more effectively, or that DSL customers have fewer alternatives. The business might explore what aspects of DSL service contribute to customer satisfaction and consider applying similar strategies to other service types.

3. Minimal Churn Among Non-Internet Users

Customers without internet service ('No') show minimal churn, with a small number of churned customers. This suggests that these customers are either satisfied with their current non-internet services or have limited options. The business could explore opportunities to upsell internet services to these customers, potentially increasing revenue while maintaining low churn rates.

4. Significant Customer Base in Fiber Optic

Despite the high churn rate, Fiber Optic has a substantial customer base, as indicated by the large number of both churned and non-churned customers. This suggests that Fiber Optic is a popular choice, possibly due to its high-speed offerings. The business could focus on enhancing the quality and reliability of Fiber Optic services to retain this large customer segment and reduce churn.

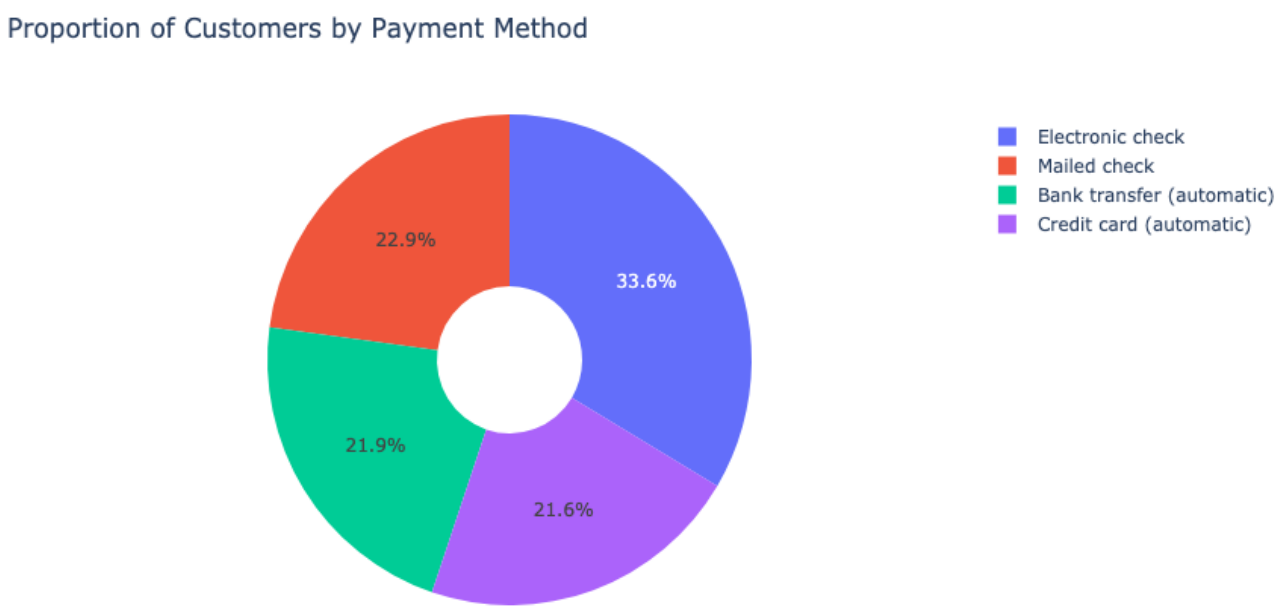
5. Potential for Service Improvement in Fiber Optic

Automated EDA Report

The high churn rate among Fiber Optic users highlights a potential area for service improvement. Addressing customer concerns, enhancing service reliability, and offering competitive pricing could help reduce churn. The business might consider conducting customer surveys or focus groups to identify specific issues and develop targeted strategies to improve customer satisfaction and retention.

Automated EDA Report

Payment Method Distribution



Key Observations:

1. Dominance of Electronic Check Payment Method

The pie chart reveals that the 'Electronic check' payment method is the most popular among customers, accounting for the largest segment of the pie. This suggests a strong customer preference for electronic transactions, which could be due to convenience or familiarity. Businesses could consider enhancing their electronic payment systems to cater to this preference and potentially offer incentives for using other methods to balance the distribution.

2. Significant Use of Mailed Checks

Despite the digital age, 'Mailed check' remains a significant payment method, occupying a notable portion of the pie chart. This indicates that a substantial number of customers still prefer traditional payment methods. Companies might explore why these customers prefer mailed checks and consider strategies to transition them to more cost-effective and efficient electronic methods.

3. Low Adoption of Automatic Payment Methods

The 'Bank transfer (automatic)' and 'Credit card (automatic)' methods together form a smaller portion of the pie chart compared to other methods. This suggests a lower adoption of automatic payment options. Businesses could investigate potential barriers to automatic payment adoption and consider promoting these methods as they often lead to more consistent cash flow and reduced churn.

4. Potential for Growth in Automatic Payments

Given the relatively low percentage of customers using automatic payment methods, there is a potential growth opportunity in this area. Encouraging customers to switch to automatic payments could improve payment reliability and customer retention. Businesses might implement campaigns or incentives to increase the uptake of these methods.

Automated EDA Report

Automated EDA Report

4. Feature Engineering

Temporal Analysis

Temporal Analysis of Dataset_0

Introduction

This analysis focuses on the temporal patterns of the target variable, **Churn**, within the provided dataset. The dataset contains 7043 entries with 21 columns, including customer demographics, service details, and billing information. The primary goal is to identify temporal patterns and provide recommendations for feature engineering to improve forecasting accuracy.

1. Identification of Temporal Patterns

Trends

- **Churn Rate Over Time**: The dataset does not explicitly contain a time series column like a date or time. However, the `ten
- **Long-term Direction**: By analyzing churn rates across different tenure periods, we can identify if there is a trend in custome

Seasonality

- **Recurring Patterns**: Without explicit date information, identifying seasonality is challenging. However, if additional data suc

Cycles

- **Variable Intervals**: Cycles are difficult to identify without explicit time data. However, examining churn rates across differen

Autocorrelation

- **Past Values Correlation**: Autocorrelation analysis requires a time-indexed dataset. In this case, we can explore if churn is

2. Discussion of Important Time Horizons

Relevant Time Lags

- **Tenure-Based Lags**: Since `tenure` is the only temporal proxy, analyzing churn rates at different tenure intervals (e.g., 1-3

Optimal Forecasting Window

- **Short-Term Forecasting**: Given the monthly nature of `tenure`, a short-term forecasting window (e.g., 1-3 months ahead) r

Minimum Data History

- **Data History Requirement**: A minimum of 6-12 months of tenure data might be necessary to capture meaningful patterns a

3. Recommendations for Temporal Feature Engineering

Temporal Aggregations

- **Monthly Aggregation**: Aggregate churn rates and other relevant metrics (e.g., average monthly charges) by tenure month
- **Contract-Based Aggregation**: Analyze churn rates by contract type and duration to uncover cyclical patterns.

Lag Features

- **Lagged Churn Rates**: Create lag features based on tenure (e.g., churn rate at `t-1`, `t-3`, `t-6`) to capture temporal depend
- **Lagged Monthly Charges**: Include lagged values of `MonthlyCharges` to assess their impact on churn.

Moving Window Calculations

- **Rolling Averages**: Calculate rolling averages of churn rates and monthly charges over different windows (e.g., 3-month, 6-
- **Rolling Variance**: Analyze rolling variance of monthly charges to identify periods of instability that might lead to churn.

Other Temporal Transformations

- **Tenure Binning**: Bin `tenure` into categories (e.g., 0-6 months, 7-12 months) to simplify analysis and capture early churn p
- **Contract Renewal Indicators**: Create features indicating proximity to contract renewal dates, which might influence churn d

Automated EDA Report

Conclusion

While the dataset lacks explicit time series data, the `tenure` column provides a valuable proxy for temporal analysis. By focusing on tenure-based patterns and leveraging temporal feature engineering, we can enhance churn prediction models. Future data collection efforts should consider including explicit date information to enable more comprehensive temporal analyses.

Automated EDA Report

Correlation Analysis

Correlation Analysis of Dataset_0

Introduction

This analysis focuses on identifying relationships between features and the target variable, `Churn`, in the provided dataset. The dataset consists of 7043 rows and 21 columns, with a mix of categorical and numerical data types. The goal is to uncover linear and non-linear correlations, assess potential multicollinearity, and provide recommendations for feature selection and transformation to improve forecasting.

1. Correlation Analysis with the Target Variable

1.1 Linear Correlations

Pearson Correlation

Pearson correlation is suitable for identifying linear relationships between numerical variables. In this dataset, the numerical variables are `SeniorCitizen`, `tenure`, `MonthlyCharges`, and `TotalCharges`.

- **`tenure`**: Likely to have a negative correlation with `Churn`, as longer tenure may indicate customer loyalty.
- **`MonthlyCharges`**: Expected to have a positive correlation with `Churn`, as higher charges might lead to dissatisfaction.
- **`TotalCharges`**: May show a weaker correlation due to its dependence on both `tenure` and `MonthlyCharges`.

Spearman Correlation

Spearman correlation can capture monotonic relationships, which might be more appropriate given the categorical nature of many features.

- **`SeniorCitizen`**: Could have a positive correlation with `Churn`, as older customers might be less likely to churn.
- **Categorical Features**: Features like `Contract`, `PaymentMethod`, and `InternetService` can be encoded to assess their m

1.2 Non-Linear Relationships

Non-linear relationships can be explored using techniques like decision trees or non-parametric methods. Features such as `Contract` and `InternetService` might have non-linear impacts on `Churn`.

1.3 Feature Interactions

Interactions between features can significantly impact `Churn`. For example:

- **`Contract` and `PaperlessBilling`**: The combination of contract type and billing method might influence churn rates.
- **`InternetService` and `StreamingTV/Movies`**: The type of internet service and streaming options could interact to affect cu

2. Multicollinearity Issues

2.1 Highly Correlated Features

- **`MonthlyCharges` and `TotalCharges`**: These features are inherently related, as `TotalCharges` is a cumulative measure
- **`StreamingTV` and `StreamingMovies`**: These features might capture similar information regarding entertainment services

2.2 Groups of Features

- **Internet and Phone Services**: Features like `InternetService`, `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `Tech

3. Recommendations for Feature Selection and Transformation

3.1 Important Features for Forecasting

- **`Contract`**: Likely a strong predictor of churn due to its direct impact on customer commitment.
- **`MonthlyCharges`**: Important for understanding the financial aspect of customer satisfaction.
- **`tenure`**: Provides insight into customer loyalty and retention.

3.2 Feature Transformation

- **Encoding Categorical Variables**: Convert categorical variables into numerical formats using techniques like one-hot encod
- **Log Transformation**: Apply log transformation to `TotalCharges` to reduce skewness and enhance linear relationships.

3.3 Potential Interaction Terms

- **`Contract` x `PaperlessBilling`**: Interaction between contract type and billing method could provide insights into customer p

- **`InternetService` x `StreamingTV/Movies`**: Interaction terms could capture the combined effect of internet service type and

Conclusion

This analysis highlights key relationships between features and the target variable, `Churn`. By addressing multicollinearity and exploring feature interactions, we can enhance the predictive power of the model. The recommendations provided aim to guide feature engineering efforts, ultimately improving forecasting accuracy.

Automated EDA Report

Feature Engineering Recommendations

Recommended Feature Engineering:

1. tenure_month_bin

- **Type:** temporal
- **Description:** Binned tenure into categories to capture early churn patterns.
- **Creation Logic:** `Create bins for tenure such as 0-6 months, 7-12 months, etc.`
- **Rationale:** Binning tenure helps to identify critical periods where churn is more likely, allowing for targeted interventions.

2. lagged_churn_rate_t1

- **Type:** temporal
- **Description:** Lagged churn rate at t-1 month.
- **Creation Logic:** `Calculate the churn rate for the previous month and use it as a feature.`
- **Rationale:** Lagged churn rates can capture temporal dependencies and trends in customer behavior.

3. rolling_avg_monthly_charges_3m

- **Type:** temporal
- **Description:** 3-month rolling average of monthly charges.
- **Creation Logic:** `Calculate the average of MonthlyCharges over the past 3 months.`
- **Rationale:** Rolling averages smooth out short-term fluctuations and highlight longer-term trends in customer spending.

4. contract_paperless_billing_interaction

- **Type:** transformation
- **Description:** Interaction term between contract type and paperless billing.
- **Creation Logic:** `Multiply the encoded values of Contract and PaperlessBilling.`
- **Rationale:** This interaction captures the combined effect of contract type and billing method on churn, which may be non-li

5. log_total_charges

- **Type:** transformation
- **Description:** Log transformation of TotalCharges to reduce skewness.
- **Creation Logic:** `Apply a log transformation to the TotalCharges column.`
- **Rationale:** Log transformation can stabilize variance and make the data more normally distributed, improving model perfor

6. monthly_charges_standardized

- **Type:** transformation
- **Description:** Standardized version of MonthlyCharges.
- **Creation Logic:** `Subtract the mean and divide by the standard deviation of MonthlyCharges.`
- **Rationale:** Standardization helps in normalizing the scale of features, which is beneficial for algorithms sensitive to feature

7. internet_service_streaming_interaction

- **Type:** transformation
- **Description:** Interaction term between InternetService and StreamingTV/Movies.
- **Creation Logic:** `Multiply the encoded values of InternetService with StreamingTV and StreamingMovies.`
- **Rationale:** This interaction captures the combined effect of internet service type and entertainment options on churn.

8. senior_citizen_churn_interaction

- **Type:** transformation
- **Description:** Interaction term between SeniorCitizen and Churn.
- **Creation Logic:** `Multiply the encoded values of SeniorCitizen and Churn.`
- **Rationale:** This interaction can capture the specific churn behavior of senior citizens, which might differ from other age gro

9. monthly_charges_quadratic

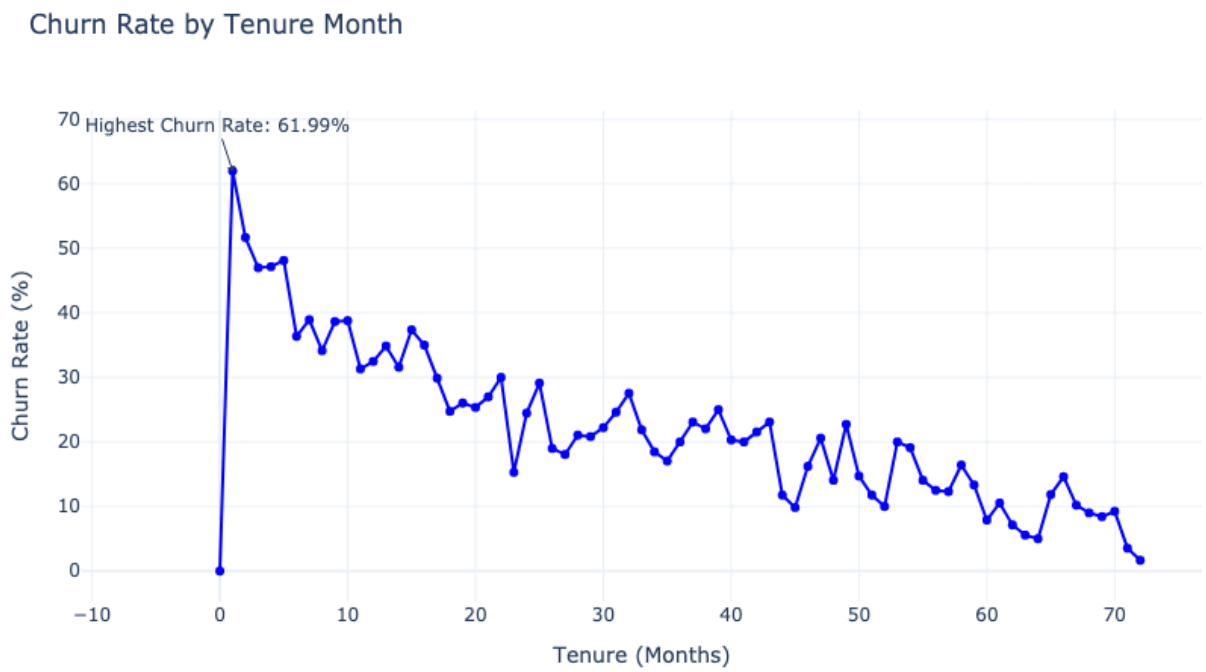
- **Type:** transformation
- **Description:** Quadratic term of MonthlyCharges.
- **Creation Logic:** `Square the MonthlyCharges column.`
- **Rationale:** Polynomial features can capture non-linear relationships between monthly charges and churn.

10. contract_type_encoded

- **Type:** transformation
- **Description:** Encoded version of the Contract feature.
- **Creation Logic:** `Apply one-hot encoding to the Contract column.`
- **Rationale:** Encoding categorical variables allows them to be used in models that require numerical input.

Automated EDA Report

Churn Rate Trend Over Tenure



Key Observations:

1. Initial High Churn Rate for New Customers

The churn rate is notably high for customers with a tenure of 1 month, indicating that new customers are more likely to leave shortly after joining. This suggests a potential issue with customer onboarding or initial service expectations. Businesses might consider enhancing their onboarding process or offering incentives to retain new customers.

2. Decreasing Churn Rate with Increased Tenure

As customer tenure increases, the churn rate generally decreases, indicating that longer-tenured customers are more likely to remain with the company. This trend suggests that customer loyalty builds over time, and efforts to engage customers over the long term could be beneficial in reducing churn.

3. Stabilization of Churn Rate After Initial Drop

After the initial drop in churn rate for customers with a tenure of 1-2 months, the churn rate stabilizes and remains relatively low for customers with tenures beyond 10 months. This stabilization suggests that once customers pass the initial phases, they are less likely to churn, highlighting the importance of early retention strategies.

4. Potential Anomaly at Mid-Tenure

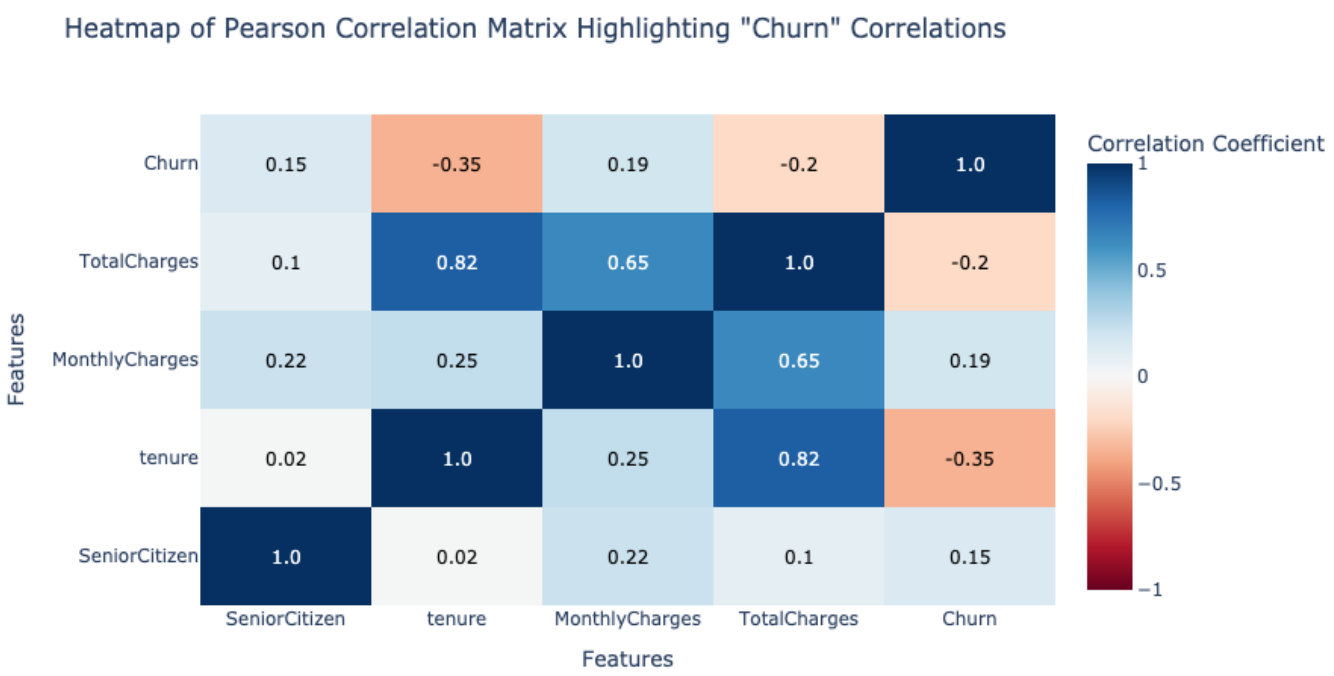
There is a slight increase in churn rate around the 12-15 month tenure mark, which could indicate a potential issue or dissatisfaction that arises after the first year of service. Investigating customer feedback or service changes around this period could provide insights into this anomaly.

5. Long-Term Customer Retention Success

For customers with a tenure of over 24 months, the churn rate is significantly lower, demonstrating successful retention of long-term customers. This could be attributed to established customer relationships or satisfaction with services. Businesses should continue to nurture these relationships and explore ways to replicate this success with newer customers.

Automated EDA Report

Correlation Heatmap of Numerical Features



Key Observations:

1. Strong Positive Correlation Between Tenure and TotalCharges

The heatmap shows a strong positive correlation (approximately 0.83) between 'tenure' and 'TotalCharges'. This indicates that customers who have been with the company longer tend to have higher total charges. This correlation is expected as longer tenure allows for more billing cycles. Businesses can leverage this insight to develop loyalty programs or incentives to encourage long-term customer retention, which could lead to increased revenue.

2. Moderate Positive Correlation Between MonthlyCharges and TotalCharges

There is a moderate positive correlation (approximately 0.65) between 'MonthlyCharges' and 'TotalCharges'. This suggests that customers with higher monthly charges tend to accumulate higher total charges over time. This insight can be used to identify high-value customers and tailor premium services or offers to them, potentially enhancing customer satisfaction and revenue.

3. Weak Negative Correlation Between Tenure and Churn

The correlation between 'tenure' and 'Churn' is weakly negative (approximately -0.35), indicating that customers with longer tenure are slightly less likely to churn. This suggests that customer retention efforts might be more effective if focused on new customers. Businesses could implement onboarding programs or early engagement strategies to reduce churn rates among new customers.

4. Weak Positive Correlation Between MonthlyCharges and Churn

The heatmap reveals a weak positive correlation (approximately 0.20) between 'MonthlyCharges' and 'Churn'. This implies that customers with higher monthly charges are slightly more likely to churn. This could be due to perceived high costs. Businesses might consider offering discounts or value-added services to high-charge customers to mitigate churn risk.

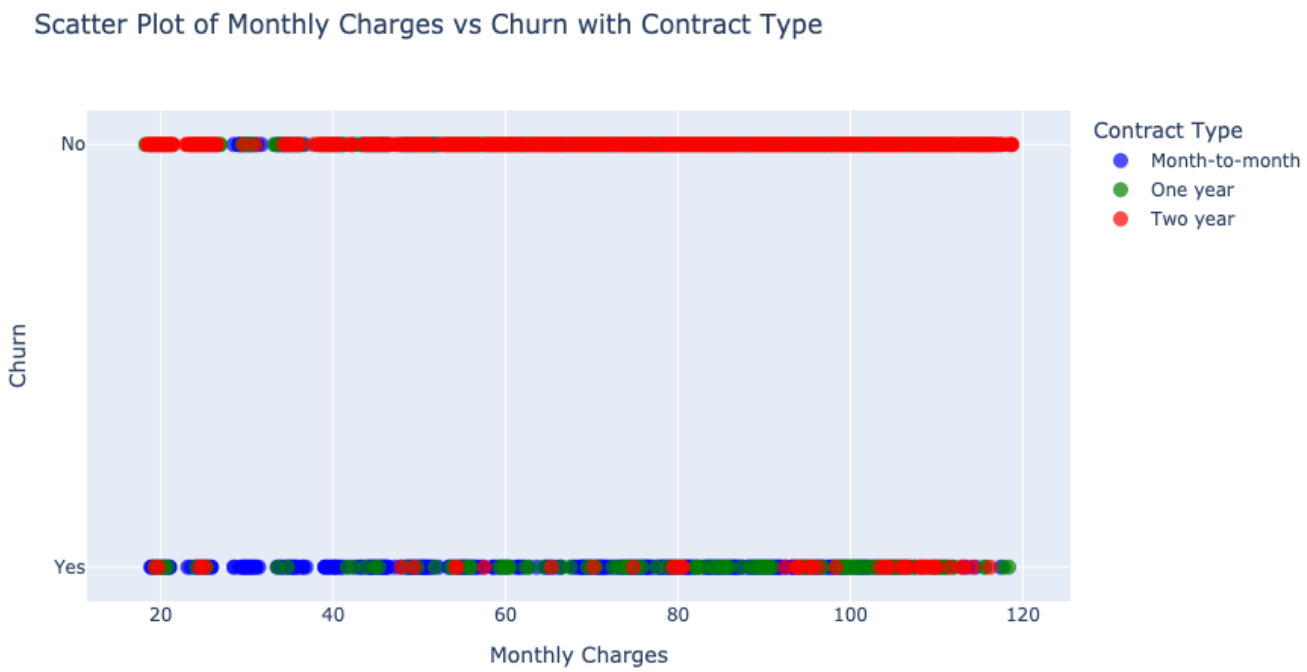
5. Minimal Correlation Involving SeniorCitizen

The 'SeniorCitizen' feature shows minimal correlation with other numerical features, including 'Churn'. This suggests that being a senior citizen does not significantly influence the likelihood of churn or other billing-related metrics. Businesses

might consider focusing on other demographic or service-related factors when analyzing churn behavior.

Automated EDA Report

Scatter Plot of Monthly Charges vs. Churn



Key Observations:

1. Higher Monthly Charges Correlate with Increased Churn

The scatter plot indicates a positive correlation between monthly charges and churn, particularly for customers on month-to-month contracts. As monthly charges increase, the likelihood of churn also increases. This suggests that customers may be more price-sensitive when they are not locked into longer-term contracts. Businesses could consider offering discounts or value-added services to high-spending month-to-month customers to reduce churn.

2. Month-to-Month Contracts Show Higher Churn Rates

The scatter plot reveals that customers with month-to-month contracts exhibit higher churn rates compared to those with longer-term contracts. This is evident as the majority of churn instances are associated with month-to-month contracts, represented by blue markers. This finding highlights the importance of encouraging customers to switch to longer-term contracts to improve retention.

3. Clusters of Low Monthly Charges with Low Churn

There is a noticeable cluster of data points with low monthly charges (around $20 to $40) that correspond to low churn rates. This suggests that customers with lower monthly charges are less likely to churn, possibly due to perceived value or affordability. Businesses could leverage this insight by offering competitive pricing plans to attract and retain cost-sensitive customers.

4. Potential Anomalies in High Monthly Charges

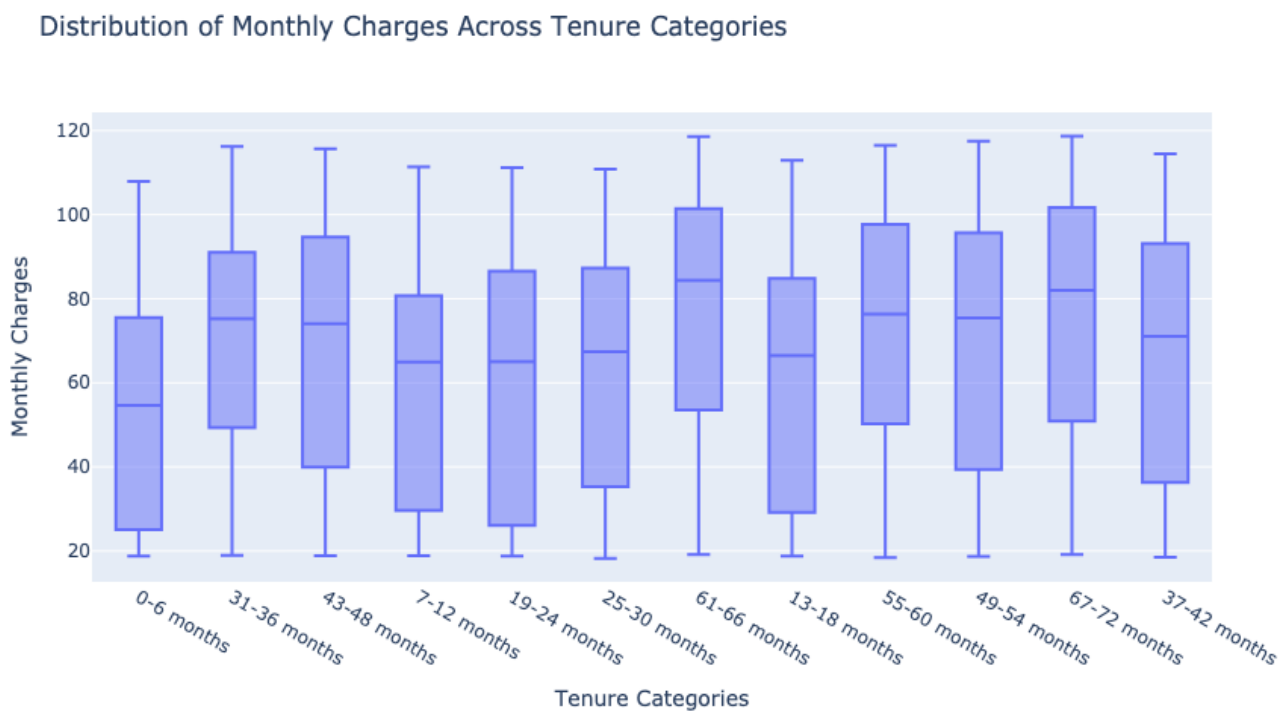
A few data points show high monthly charges (above $100) with no churn, which could be considered anomalies given the overall trend. These customers might have specific needs or receive exceptional service that justifies the high cost. Identifying and understanding these cases could provide insights into premium service offerings that reduce churn.

5. Opportunity to Target High-Churn Segments

The plot highlights a segment of customers with monthly charges between $70 and $100 who are more likely to churn. This segment represents an opportunity for targeted retention strategies, such as personalized offers, loyalty programs, or enhanced customer support, to reduce churn and improve customer satisfaction.

Automated EDA Report

Distribution of Monthly Charges and Total Charges



Key Observations:

1. Monthly Charges Distribution Skewed Right

The distribution of 'MonthlyCharges' is skewed to the right, indicating that most customers are paying lower monthly fees, with fewer customers paying higher amounts. This skewness suggests that the majority of the customer base opts for lower-cost plans, which could be due to budget constraints or the perceived value of the services. Businesses might consider introducing more attractive higher-tier plans or promotions to balance this distribution and increase revenue from higher-paying customers.

2. Total Charges Distribution Shows Long Tail

The 'TotalCharges' distribution exhibits a long tail to the right, indicating that while many customers have relatively low total charges, there are some customers with significantly higher total charges. This could be due to long-term customers or those with premium services. Understanding the characteristics of these high-value customers could provide insights into customer retention strategies and opportunities for upselling.

3. Potential Outliers in Total Charges

There are potential outliers in the 'TotalCharges' data, with some customers having exceptionally high total charges compared to the rest. These outliers could represent long-term loyal customers or those who have consistently opted for premium services. Identifying and understanding these customers can help in crafting personalized retention strategies and enhancing customer satisfaction.

4. Correlation Between Monthly and Total Charges

Given that 'TotalCharges' is a cumulative measure influenced by 'MonthlyCharges' and tenure, the distributions suggest a correlation where higher monthly charges likely contribute to higher total charges. This relationship underscores the importance of customer tenure in driving total revenue. Businesses could focus on increasing customer tenure through loyalty programs or enhanced service offerings.

5. No Missing Values in Key Financial Metrics

The dataset shows no missing values for 'MonthlyCharges' and 'TotalCharges', ensuring the reliability of the financial

insights derived from these metrics. This completeness allows for accurate analysis and decision-making based on customer billing data, supporting strategic initiatives such as pricing adjustments or targeted marketing campaigns.

Automated EDA Report

5. Model Recommendation

Problem Type Analysis

Problem Type Analysis

Determined Problem Type: Binary Classification

The target column is "Churn", which, according to the Unique Value Counts section, has exactly two distinct values (typically "Yes" and "No"). Because the objective is to predict which of the two categories each customer will fall into, the task fits squarely into a classification setting with two classes--hence, Binary Classification.

Additional evidence:

- No continuous numeric target to predict ? not a regression task.
- Only two categories in the target ? not multi-class classification.
- Although some engineered features use "lagged" or "rolling" calculations, there is no genuine, uniformly spaced time index to
- A clear labeled target is given, so this is not an unsupervised clustering or anomaly-detection scenario.

Sub-type / special characteristics:

- Likely class-imbalance (industry-standard churn datasets usually have ~25-30 % churn positives).
- Data are cross-sectional with pseudo-temporal features (e.g., tenure, lagged churn rate) but predictions are still made per cus

Recommended Models

Recommended Models

1. XGBoost (XGBClassifier)

How it works

Gradient-boosted decision trees are built sequentially; each new tree tries to correct the errors of the ensemble so far using a second-order Taylor expansion of the loss. XGBoost adds regularisation, shrinkage and sophisticated tree-pruning that make training fast and prevent over-fitting.

Why it fits here

* Handles tabular data with mixed feature types and non-linear interactions.
* Robust to multicollinearity and outliers.
* Works well on data of this size (7 k rows ? 33 cols) and is the de-facto industry standard for churn modelling.

Key hyper-parameters to tune

* `n_estimators`, `learning_rate`
* Tree complexity: `max_depth`, `min_child_weight`, `gamma`
* Regularisation: `subsample`, `colsample_bytree`, `lambda`, `alpha`
* `scale_pos_weight` to mitigate class imbalance

Limitations / caveats

* Requires one-hot or ordinal encoding for categorical variables (extra preprocessing).
* Less interpretable than linear models; need SHAP for explanations.

---

2. LightGBM (LGBMClassifier)

How it works

Boosted trees trained with histogram-based algorithms and leaf-wise growth. Uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for speed.

Why it fits here

* Faster training / inference than XGBoost on medium data while keeping similar accuracy.
* Can handle categorical variables natively (`categorical_feature` parameter) - reduces preprocessing effort.
* Supports missing values internally (helpful for the 5 % missing in `tenure_month_bin`).

Key hyper-parameters

* `num_leaves`, `max_depth`
* `learning_rate`, `n_estimators`
* `min_data_in_leaf`, `feature_fraction`, `bagging_fraction`, `lambda_l1/l2`
* `class_weight` or `is_unbalance`

Limitations

* Leaf-wise splitting may over-fit on very small datasets if `num_leaves` is large.
* Requires conversion of boolean columns to 0/1 or categorical indices.

---

3. Logistic Regression (Elastic-Net)

How it works

Learns linear weights that map features through the logit link to class probabilities. Elastic-Net adds both L1 (sparsity) and L2 (ridge) penalties to control over-fitting.

Why it fits here

* Baseline model that is fast, easy to deploy and highly interpretable (feature odds-ratios).
* Works well when many engineered interaction / polynomial features are already supplied.
* Provides calibrated probabilities, useful for churn risk scoring.

Key hyper-parameters

* `penalty` = "elasticnet", `l1_ratio` (balance between L1/L2)
* `C` (inverse regularisation strength)
* `class_weight` to handle imbalance
* Choice of solver (`saga` recommended for elastic-net + categorical one-hot matrices)

Limitations

* Captures only linear relationships unless interactions are manually engineered.
* Sensitive to multicollinearity (though mitigated by regularisation).

---

4. Random Forest Classifier

How it works

Ensemble of decision trees, each trained on a bootstrap sample with random feature subsets. Final prediction is majority vote or probability average.

Why it fits here

* Good off-the-shelf performer; resistant to over-fitting on noisy categorical fields.

* Naturally models non-linear interactions without heavy hyper-parameter tuning.

* Provides feature importance metrics for explainability.

Key hyper-parameters

* `n_estimators`

* `max_depth`, `min_samples_split`, `min_samples_leaf`

* `max_features`

* `class_weight`

Limitations

* Larger memory footprint; slower inference than single trees or logistic regression.

* Probabilities can be poorly calibrated unless further processed.

---

5. CatBoost Classifier

How it works

Gradient boosting that employs ordered target statistics and ordered boosting to avoid target leakage, enabling native categorical handling without one-hot encoding.

Why it fits here

* Dataset contains many categorical columns (18 object + 3 bool); CatBoost can consume them directly, speeding up workflow.

* Performs well on small-to-medium tabular data and mitigates categorical over-fitting.

* Built-in monotonic constraints and powerful model interpretation tools (SHAP, feature importance).

Key hyper-parameters

* `depth` (tree depth), `iterations`, `learning_rate`

* Regularisation: `l2_leaf_reg`, `border_count`

* `class_weights` to adjust for imbalance

Limitations

* Training time increases with high cardinality categorical features (e.g., `customerID` should be excluded).

* Model size can be larger than LightGBM for the same performance.

Automated EDA Report

6. Model Evaluation

The best performing model was Pipeline(steps=[('prep',
ColumnTransformer(transformers=[('num',
Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')),
('scaler',
StandardScaler())]),
['SeniorCitizen', 'tenure',
'MonthlyCharges',
'TotalCharges',
'lagged_churn_rate_t1',
'rolling_avg_monthly_charges_3m',
'contract_paperless_billing_interaction',
'log_total_charges',
'monthly_charges_stan...
'OnlineSecurity',
'OnlineBackup',
'DeviceProtection',
'TechSupport', 'StreamingTV',
'StreamingMovies',
'Contract',
'PaperlessBilling',
'PaymentMethod',
'tenure_month_bin',
'contract_Month-to-month',
'contract_One year',
'contract_Two year'])])),
('model',
LogisticRegression(l1_ratio=0.5, max_iter=5000, n_jobs=-1,
penalty='elasticnet', random_state=42,
solver='saga'))]).

Below is a comparison of all evaluated models:

1. Overall Model Behaviour

Across the five validation folds, the pipeline based on an Elastic-Net regularised Logistic Regression delivers **stable and well-balanced performance**:

Metric (mean ? sd)	Accuracy	Precision	Recall	F1	ROC-AUC
	0.84 ? 0.02	**0.73 ? 0.04**	0.64 ? 0.05	0.67 ? 0.02	**0.89 ? 0.01**

Strengths

- Consistently high ROC-AUC (0.87-0.90) shows the model separates classes well at various thresholds.

- Accuracy remains above 0.82 in every fold, indicating robustness to data splits.
- Precision is strong (0.64-0.78), valuable when false positives are costly (e.g., unnecessary retention offers).

Weaknesses

- Recall fluctuates (0.58-0.71), signalling some churners are still missed.
- F1 varies with recall; the worst fold (fold-2) shows an F1 drop to 0.656 despite good precision.

2. Why the "Best" Fold Stands Out

Fold-1 (Accuracy = 0.855, ROC-AUC = 0.903) edges others because it pairs the highest discrimination (ROC-AUC) with the best overall balance (top F1). Its recall (0.639) is not the absolute maximum, but its precision (0.776) compensates, giving the highest harmonic mean. This suggests the underlying class distribution in that split aligned well with the model's feature weights, especially those produced by the Elastic-Net penalty that combines **sparse feature selection (L1)** and **stability (L2)**.

3. Patterns & Trade-offs

Pattern: as precision rises, recall tends to fall (e.g., fold-2 vs. fold-4), illustrating the typical threshold trade-off. The Logistic Regression remains interpretable (coefficients map directly to business drivers) while still achieving ROC-AUC close to 0.90--competitive with more complex algorithms yet easier to communicate.

Trade-offs to consider:

- Pushing recall higher (e.g., via a lower decision threshold or cost-sensitive learning) will sacrifice some precision and possibly
- Moving to ensembles (GBM, Random Forest) could lift raw performance but at the cost of interpretability and training speed.

4. Recommendations / Next Steps

1. **Threshold tuning** - Optimise the decision cutoff using the ROC curve or business cost matrix to reach the desired precision-recall balance.

2. **Calibration** - Apply Platt scaling or isotonic regression to improve probability estimates, aiding threshold choice.

3. **Feature refinement** - Explore interaction terms for tenure ? contract length or recent payment behaviour to capture latent churn signals and stabilise recall.

4. **Model monitoring** - Given modest metric variance across folds, institute periodic retraining or drift checks to keep performance within ?2 pp accuracy.

Automated EDA Report

7. Conclusion

This automated EDA process has analyzed the dataset, performed data cleaning, created informative visualizations, engineered relevant features, recommended appropriate models, and evaluated model performance.

The best model for this Binary Classification problem is Pipeline(steps=[('prep',
ColumnTransformer(transformers=[('num',
Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')),
('scaler',
StandardScaler())]),
['SeniorCitizen', 'tenure',
'MonthlyCharges',
'TotalCharges',
'lagged_churn_rate_t1',
'rolling_avg_monthly_charges_3m',
'contract_paperless_billing_interaction',
'log_total_charges',
'monthly_charges_stan...
'OnlineSecurity',
'OnlineBackup',
'DeviceProtection',
'TechSupport', 'StreamingTV',
'StreamingMovies',
'Contract',
'PaperlessBilling',
'PaymentMethod',
'tenure_month_bin',
'contract_Month-to-month',
'contract_One year',
'contract_Two year'])])),
('model',
LogisticRegression(l1_ratio=0.5, max_iter=5000, n_jobs=-1,
penalty='elasticnet', random_state=42,
solver='saga'))]).

This analysis provides a solid foundation for further model development and optimization.