Telecom Customers Churn

Abdelghafor's Hackathon

Prepared by Basmala Salama , Mohamed Hany and Ziad Henedy

Overview

Goals:

- Analyzing customer behavior in the telecom industry.
- Build predictive models that classify customers into churn or non-churn categories.
- Segment customers into groups based on shared characteristics.

Agenda:

Data Understanding

- Problem Statement
- Data Dictionary
- EDA

Data Preprocessing

- Handling missing data
- Encoding and Scaling
- Resampling
- Feature Selection
- Dimensionality Reduction

Supervised Learning Model

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost

Unsupervised Learning Model

- K-Means Clustering
- Hierarchical Clustering

Overall Analysis & Insights

- Business Insights
- Recommendation

Problem Statement

The Telco customer churn data contains information about a telecom company named Telco that provide home phone and internet services to 7043 customers. Telco has observed a decline in customer retention rates so it provides this data which indicates which customers have left, stayed, or signed up for their service. Multiple important demographics and services are included for each customer, totaling 20 features. **The company wants to identify the factors contributing to customer churn, develop strategies to improve customer retention and tailor services to different customer segments.**

Data Dictionary

Churn	***Yes*** **= the customer left the company within the last month.** ***No*** **= the customer remained with the company.**
gender	customer's gender: ***Male***, ***Female***
SeniorCitizen	customer is 65 or older: ***1***, ***0*** **(meaning** ***Yes*** **and** ***No*****, respectively)**
Partner	customer is married: ***Yes***, ***No***
Dependents	customer lives with any dependents: ***Yes***, ***No*****. Dependents could be children, parents, grandparents, etc.**
PhoneService	customer subscribes to home phone service with the company: ***Yes***, ***No***
MultipleLines	customer subscribes to multiple telephone lines with the company: ***Yes***, ***No***, ***No internet service***
InternetService	customer subscribes to Internet service with the company: ***No***, ***DSL***, ***Fiber Optic***
OnlineSecurity	customer subscribes to an additional online security service provided by the company: ***Yes***, ***No***, ***No internet service***
OnlineBackup	customer subscribes to an additional online backup service provided by the company: Yes, No, No internet service
DeviceProtection	customer subscribes to an additional device protection plan for their Internet equipment provided by the company: ***Yes***, ***No***, ***No internet service***
TechSupport	customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No, No internet service
StreamingTV	**customer uses their Internet service to stream television programming from a third-party provider: Yes, No, No internet service**
StreamingMovies	customer uses their Internet service to stream movies from a third-party provider: Yes, No, No internet service
tenure	total number of months that the customer has been with the company.
Contract	customer's current contract type: Month-to-Month, One Year, Two Year.
PaperlessBilling	customer has chosen paperless billing: Yes, No
PaymentMethod	how the customer pays their bill: Electronic check, Credit Card, Mailed Check, Bank transfer
MonthlyCharge	customer's current total monthly charge for all their services from the company
TotalCharges	customer's total charges, calculated to the end of the quarter
CustomerID	unique ID that identifies the customer

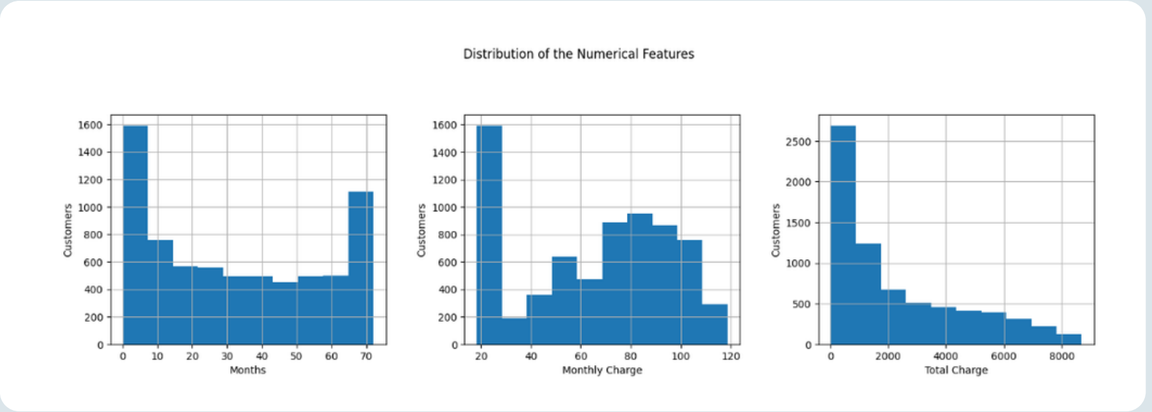
Exploratory Data Analysis

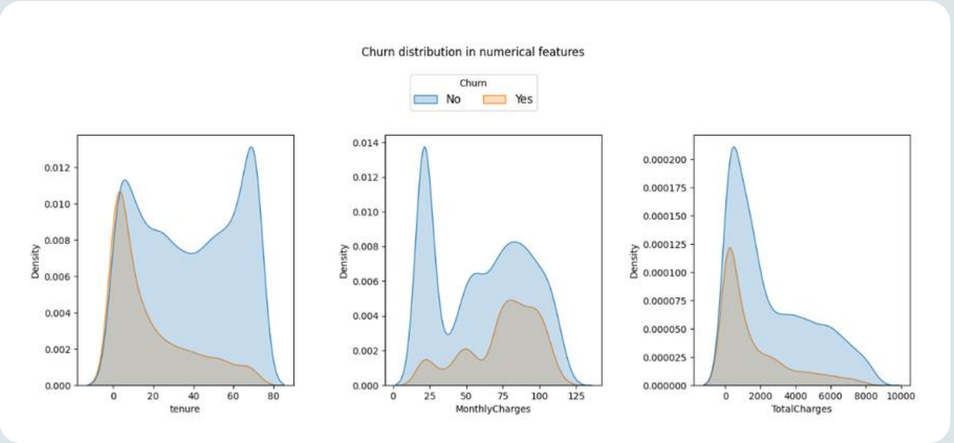
The tenure distribution has an interesting shape. Most customers have been with the company for just a few months, but also many have been for about 72 months (maximum value for tenure). This is probably related with different contracts, something that we will check soon. Probably some marketing campaign was ran recently to capture new customers due to the high number of customers with few months.
We can see that most customers pay low monthly charges, but there is a great fraction with medium values. Since most customers have been with the company for just a few months, the total charges plot shows most customers with low values.

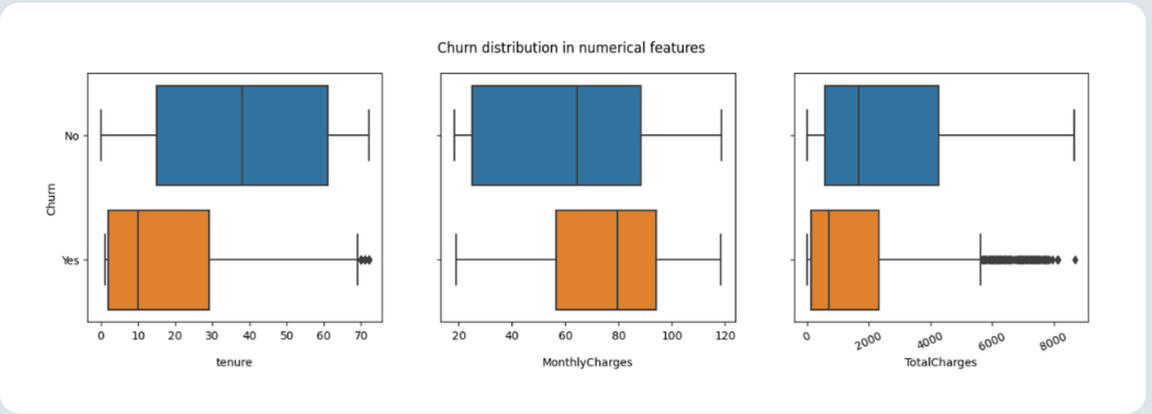
The plot shows that short tenure (recent) customers have higher churn rates. Moreover, the higher the monthly charge, the higher the churn rate.

The boxplots show that the churn rate is higher among customers with low tenure and high monthly charges. In details:

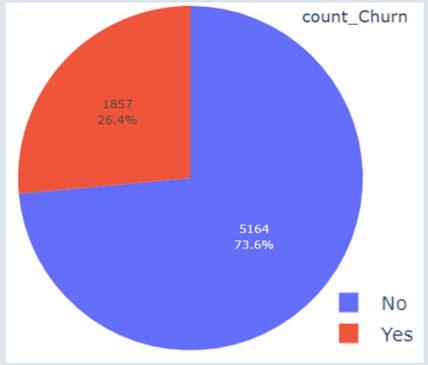
- the median tenure for customers who have left is around 10 months, while it is around 40 months for those who have stayed with the company
- the median monthly charge for customers who have churned is around 80, while it is around 65 for those who have not churned
- since most customers who have churned spent less time with the company, they have low total charges compared with those who have stayed
- There are many outliers in the total charges boxplot of customers who have churned. It is not clear the cause, but it could be wrong billing or expensive services that guided the customers away from the company.

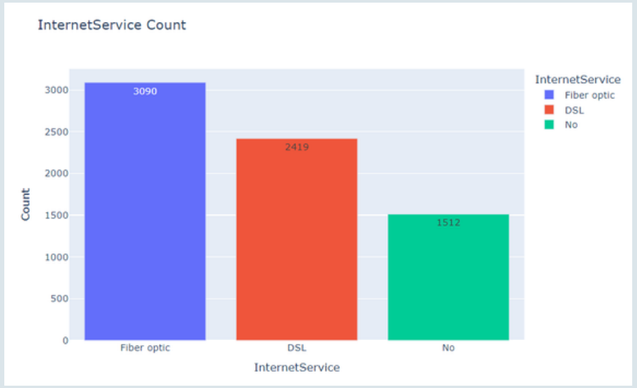


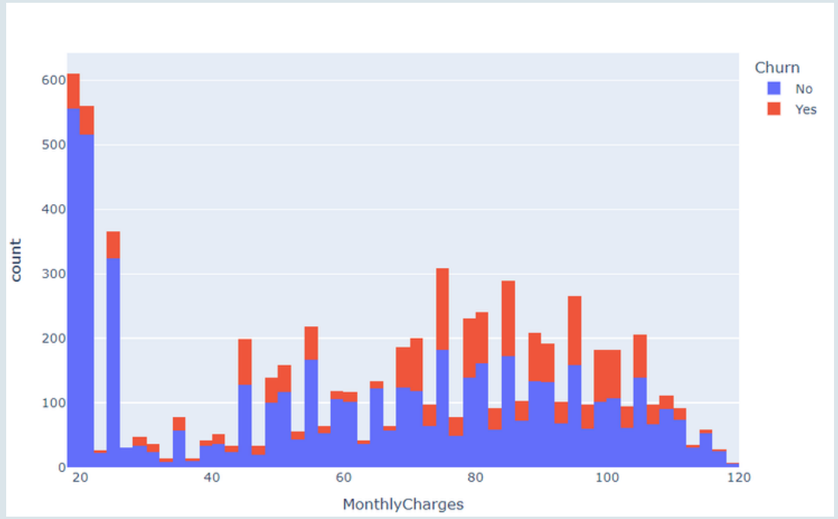


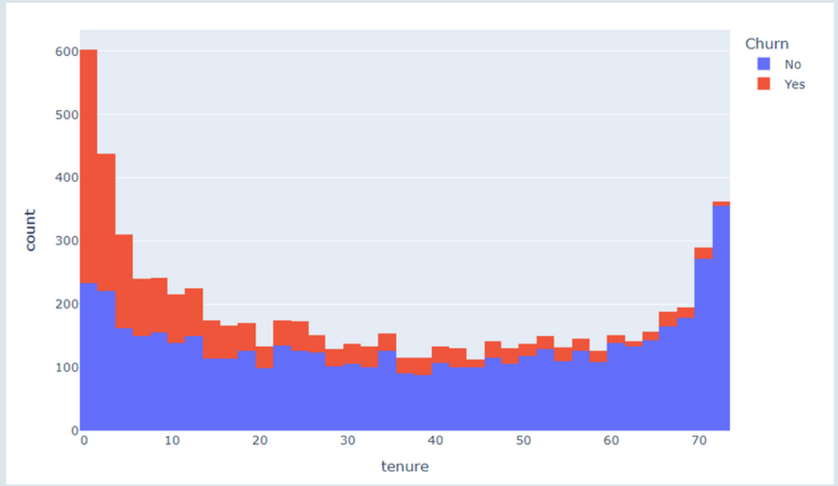


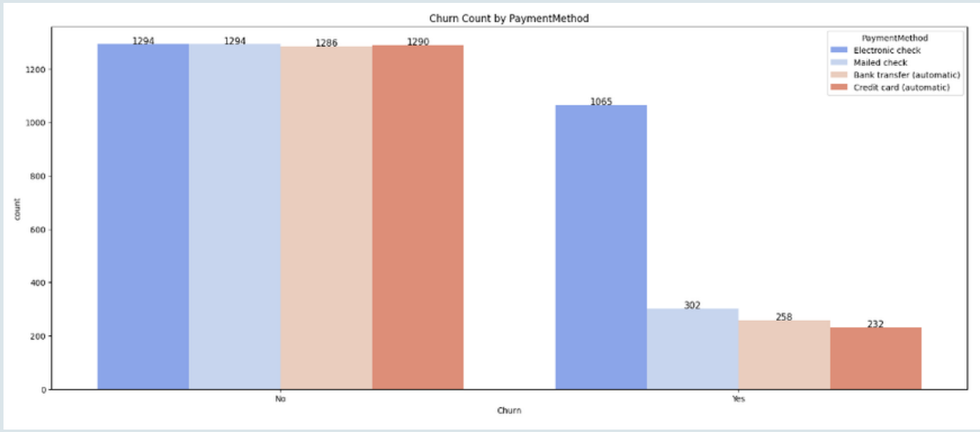
Exploratory Data Analysis

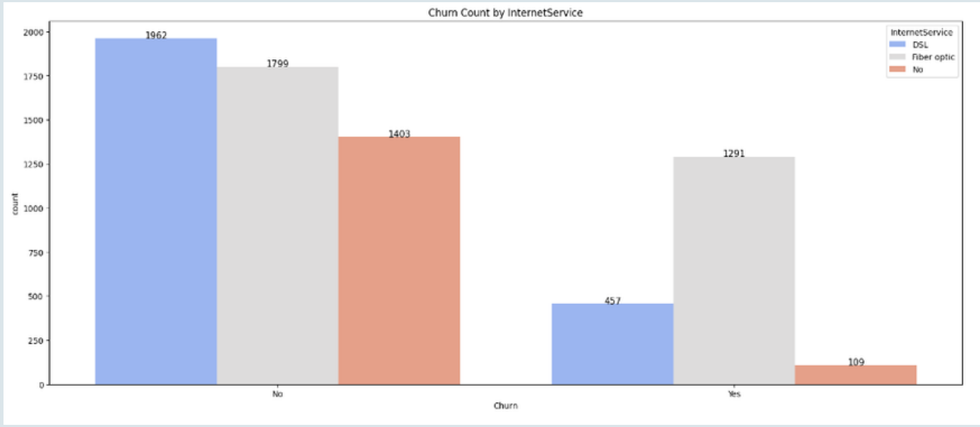


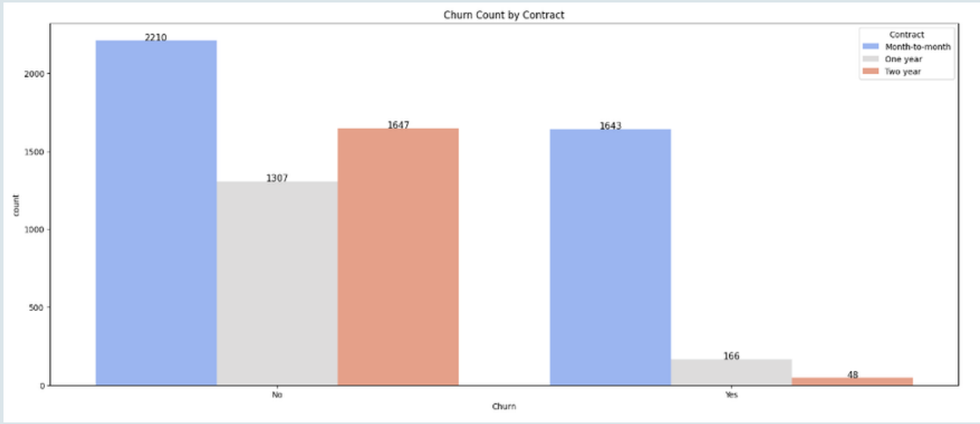












Data Preprocessing

Handling missing data

- We start by removing customerID which is unnecessary column, handling missing values in TotalCharges, and converting data types where necessary.
- After replacing some string values like 'No internet service' and 'No phone service' with 'No', we handle the empty TotalCharges by replacing them with the median.

Encoding and Scaling

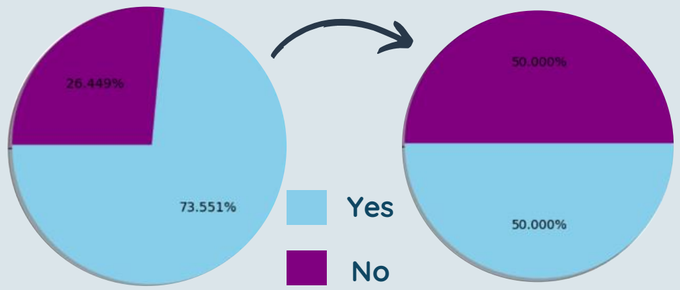
- Encoding Categorical Variables

1. **Label Encoding:** Replaced categorical binary variables with numerical equivalents.
2. **One-Hot Encoding:** Applied one-hot encoding to multi-category variables.

- **Min-Max Scaling:** Scaled MonthlyCharges, TotalCharges, and tenure to a range between 0 and 1.

Resampling

- After identifying class imbalance in the target variable (Churn), you employ **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the dataset.



Data Preprocessing

Feature Selection

- **Pearson Correlation Analysis:** We computed the correlation matrix to identify highly correlated features, which can lead to multicollinearity issues. Based on the correlation matrix, you decide which features to drop TotalCharges and InternetService_Fiber optic due to high correlation with other features.
- **Fisher Score (ANOVA F-value):** we calculate the Fisher Score to rank features based on their importance in predicting churn. This score helps to select the top 12 features for model building.

Dimensionality Reduction

- **Principal Component Analysis (PCA):** is performed to reduce dimensionality, explaining how the components capture variance in the data to determine how many components to retain for analysis.
- **Linear Discriminant Analysis (LDA):** We explored LDA as a dimensionality reduction technique. We decided not to use LDA because the variables were not statistically independent, and the classes could not be adequately separated using linear combinations.

Supervised Learning Model

Logistic Regression

- Purpose: Predicts binary outcomes (e.g., churn vs. non-churn).
- Key Advantage: Simple and interpretable.

Decision Trees

- Purpose: Models decisions with a tree-like structure.
- Key Advantage: Intuitive and easy to visualize.

Random Forest

- Purpose: Combines multiple decision trees for better accuracy.
- Key Advantage: Reduces overfitting and improves robustness.

XGBoost

- Purpose: Builds strong models by correcting previous errors.
- Key Advantage: Fast, efficient, and highly accurate.

Logistic Regression

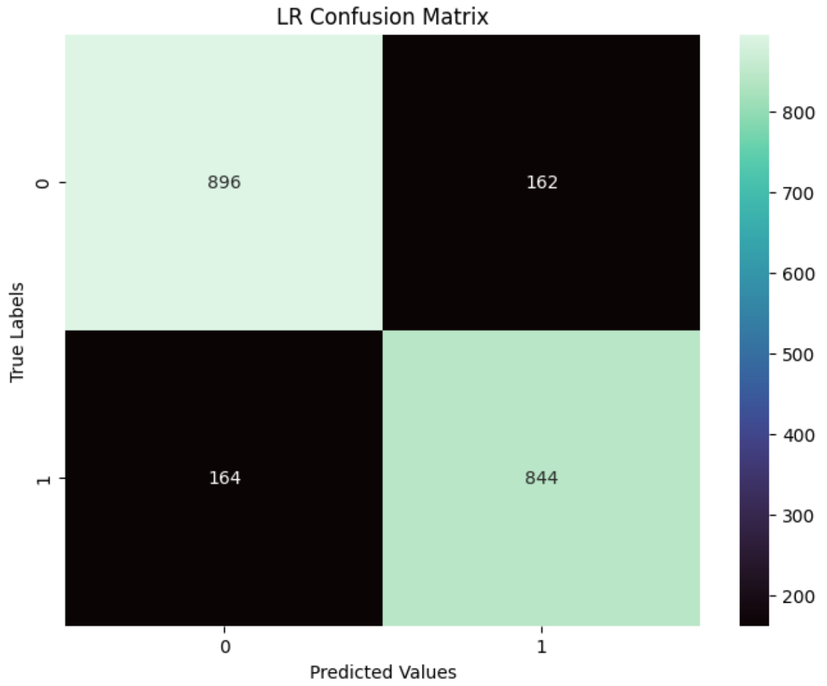
We began by training a Logistic Regression model without penalties to establish a baseline, followed by generating training and testing reports to evaluate performance. A confusion matrix was used to visualize classification results. Next, we performed hyperparameter tuning using Randomized Search to optimize key parameters such as regularization and penalty type. After identifying the best parameters, we trained the final Logistic Regression model and generated new performance reports, followed by an updated confusion matrix to visualize results.

Training Report

	precision	recall	f1-score	support
0	0.8368	0.8490	0.8428	4106
1	0.8486	0.8364	0.8425	4156
accuracy			0.8427	8262
macro avg	0.8427	0.8427	0.8427	8262
weighted avg	0.8427	0.8427	0.8427	8262

Testing Report

	precision	recall	f1-score	support
0	0.8453	0.8469	0.8461	1058
1	0.8390	0.8373	0.8381	1008
accuracy			0.8422	2066
macro avg	0.8421	0.8421	0.8421	2066
weighted avg	0.8422	0.8422	0.8422	2066



Decision Trees

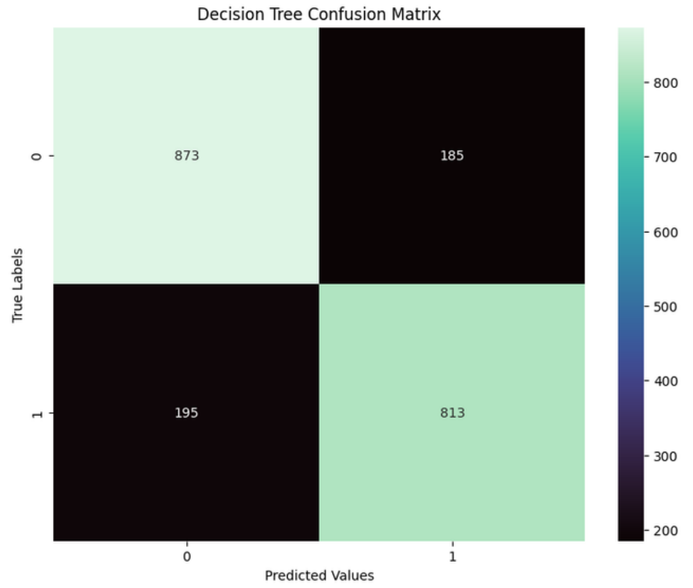
We first trained a Decision Tree model on the training data and generated training and testing reports to evaluate performance, observing overfitting. A confusion matrix visualized classification results. Next, we performed hyperparameter tuning using Randomized Search and Optuna to optimize parameters like min_samples_split, max_depth, and max_features. The tuned model was retrained, and new reports and a confusion matrix were generated. We also extracted the top 10 most important features identified by the model after training, such as customer attributes or behaviors relevant to churn prediction

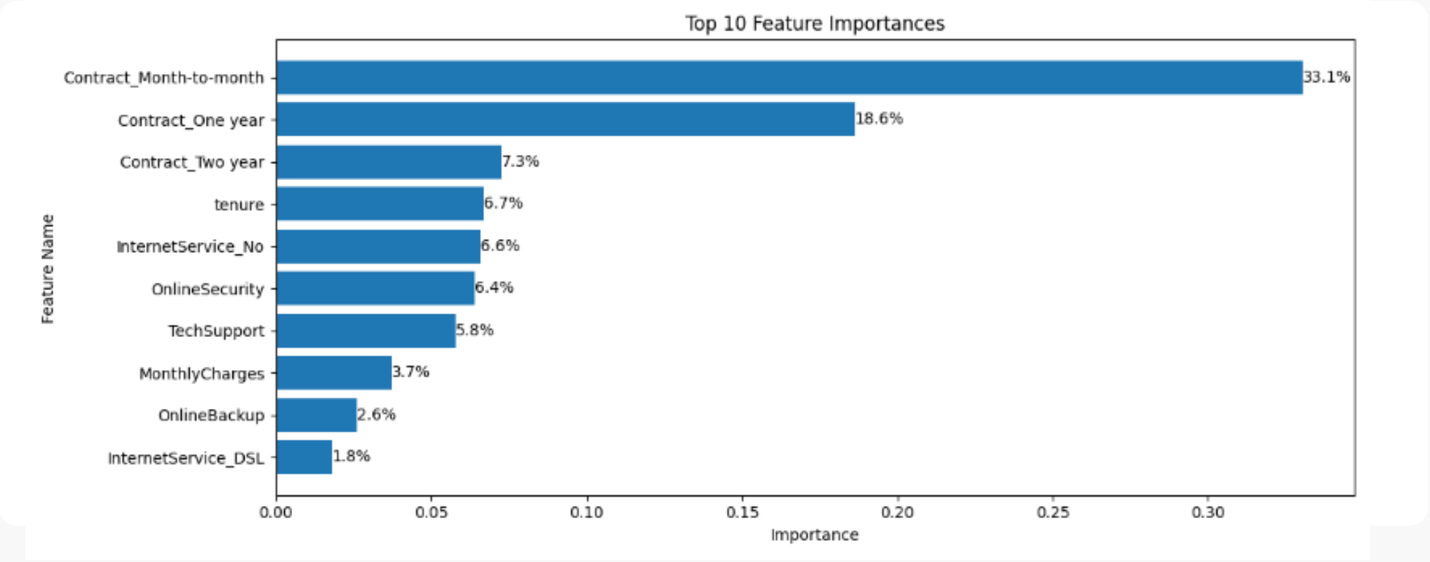
Training Report

	precision	recall	f1-score	support
0	0.8312	0.8419	0.8365	4106
1	0.8418	0.8311	0.8364	4156
accuracy			0.8365	8262
macro avg	0.8365	0.8365	0.8365	8262
weighted avg	0.8365	0.8365	0.8365	8262

Testing Report

	precision	recall	f1-score	support
0	0.8174	0.8251	0.8213	1058
1	0.8146	0.8065	0.8106	1008
accuracy			0.8161	2066
macro avg	0.8160	0.8158	0.8159	2066
weighted avg	0.8161	0.8161	0.8160	2066





Random Forest

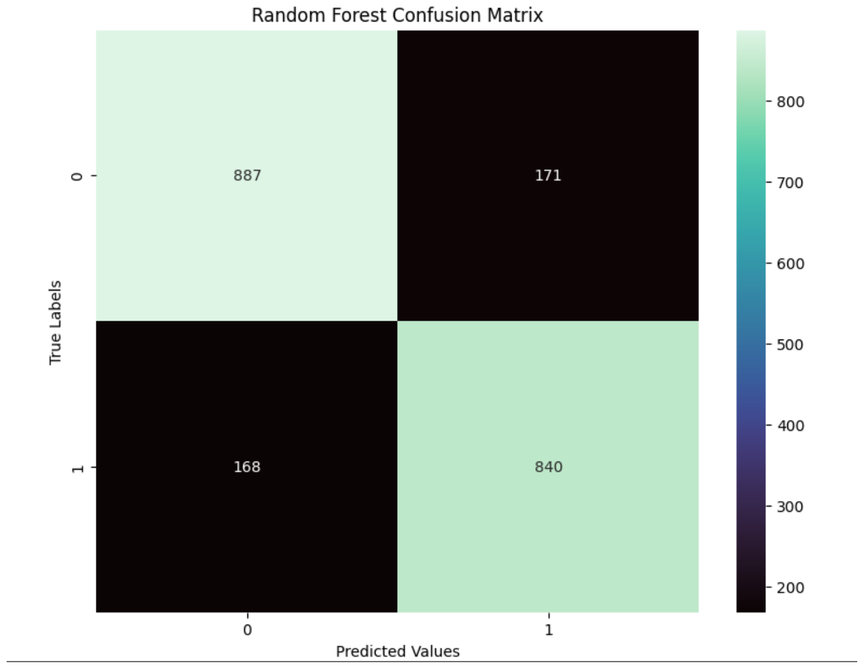
We first trained a base Random Forest model on the training data and generated training and testing reports to evaluate its performance. A confusion matrix was used to visualize the classification results. Next, we tuned hyperparameters using Randomized Search and Grid Search to optimize values like min_samples_split, max_depth, and max_features for better regularization and complexity control. After identifying the best parameters, we retrained the Random Forest model, generated new performance reports, and created an updated confusion matrix to visualize the classification results.

Training Report

	precision	recall	f1-score	support
0	0.8824	0.8843	0.8833	4106
1	0.8855	0.8835	0.8845	4156
accuracy			0.8839	8262
macro avg	0.8839	0.8839	0.8839	8262
weighted avg	0.8839	0.8839	0.8839	8262

Testing Report

	precision	recall	f1-score	support
0	0.8408	0.8384	0.8396	1058
1	0.8309	0.8333	0.8321	1008
accuracy			0.8359	2066
macro avg	0.8358	0.8359	0.8358	2066
weighted avg	0.8359	0.8359	0.8359	2066



XGBoost

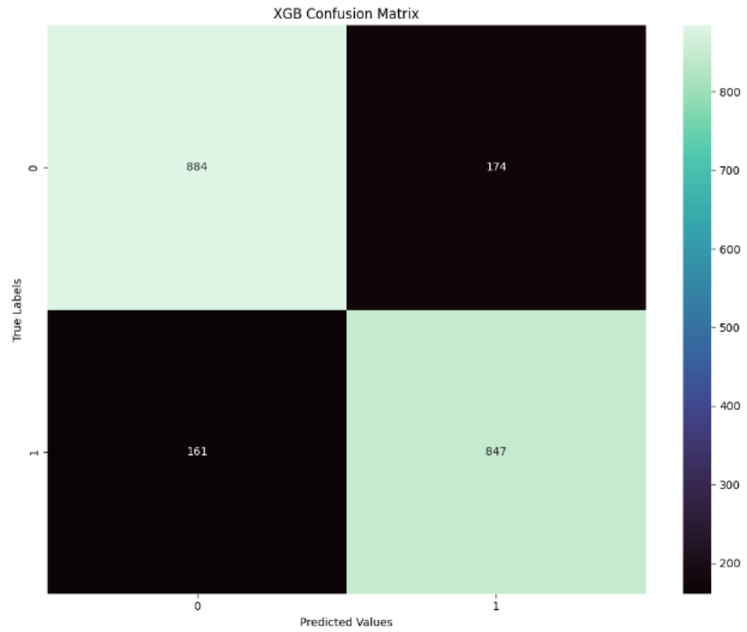
We began by training a base XGBoost classifier, generating training and testing reports to evaluate its performance, and visualizing the results with a confusion matrix. After identifying the top five important features, we removed the least important one from both the training and testing datasets. We then tuned hyperparameters using Randomized Search and Grid Search to optimize parameters such as max_depth, learning_rate, and reg_lambda. After retraining the model with the best parameters, we assessed its performance again with new classification reports and a confusion matrix. Further enhancement was achieved through hyperparameter tuning with Optuna, leading to additional evaluations. Ultimately, XGBoost was selected as the final model due to its superior precision, recall, and F1-score, demonstrating the best overall performance.

Training Report

	precision	recall	f1-score	support
0	0.9428	0.9313	0.9370	4106
1	0.9330	0.9442	0.9385	4156
accuracy			0.9378	8262
macro avg	0.9379	0.9377	0.9378	8262
weighted avg	0.9378	0.9378	0.9378	8262

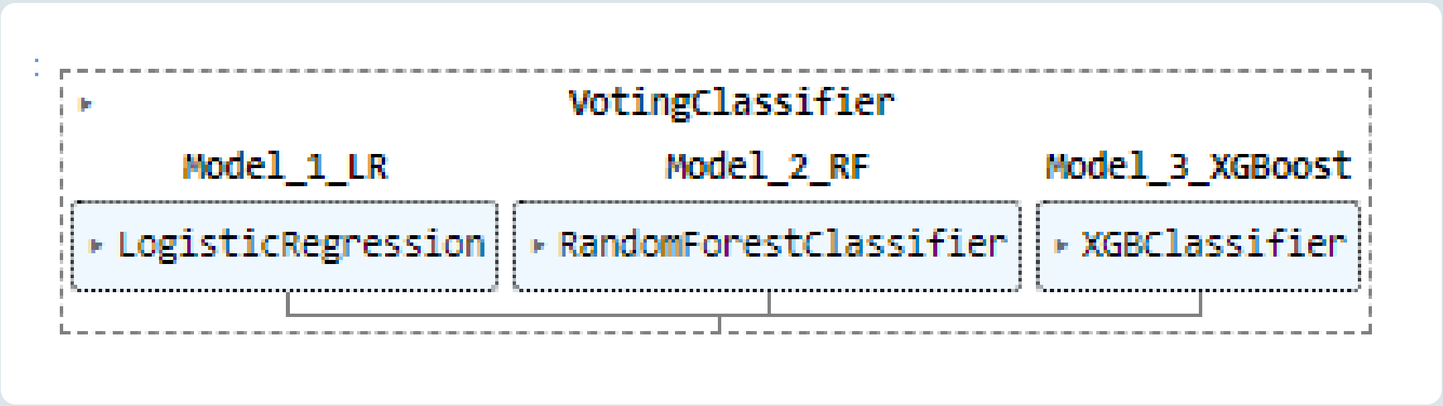
Testing Report

	precision	recall	f1-score	support
0	0.8459	0.8355	0.8407	1058
1	0.8296	0.8403	0.8349	1008
accuracy			0.8379	2066
macro avg	0.8378	0.8379	0.8378	2066
weighted avg	0.8380	0.8379	0.8379	2066



Voting Classifier

To enhance our model's performance, we implemented a Voting Classifier that combined multiple base models, including XGBoost, Random Forest, and Logistic Regression. This ensemble approach aimed to leverage the strengths of each individual model, potentially improving overall accuracy and robustness. After training the Voting Classifier, we evaluated its performance through classification reports and confusion matrices. The results showed that the Voting Classifier achieved comparable performance to XGBoost, effectively combining the predictions of its constituent models. This strategy demonstrated the potential benefits of using ensemble methods, leading to reliable predictions across the different classes.

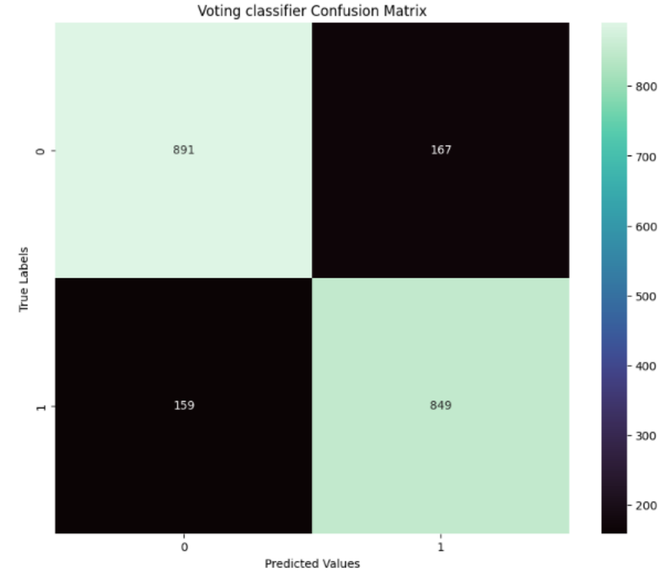


Training Report

	precision	recall	f1-score	support
0	0.8972	0.8926	0.8949	4106
1	0.8944	0.8989	0.8967	4156
accuracy			0.8958	8262
macro avg	0.8958	0.8958	0.8958	8262
weighted avg	0.8958	0.8958	0.8958	8262

Testing Report

	precision	recall	f1-score	support
0	0.8486	0.8422	0.8454	1058
1	0.8356	0.8423	0.8389	1008
accuracy			0.8422	2066
macro avg	0.8421	0.8422	0.8421	2066
weighted avg	0.8423	0.8422	0.8422	2066



Unsupervised Learning Model

K-Means Clustering

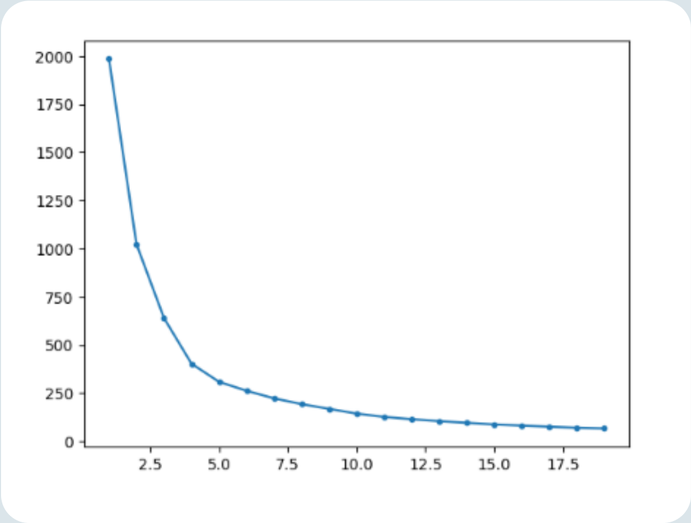
- **Purpose:** Groups data points into a predefined number (K) of clusters, where each point belongs to the cluster with the nearest mean.
- **How it Works:** It assigns data points to K clusters based on their distance to the cluster centers (centroids), then iteratively adjusts the centroids until the clusters stabilize.
- **Key Advantage:** Simple to implement and computationally efficient for large datasets.

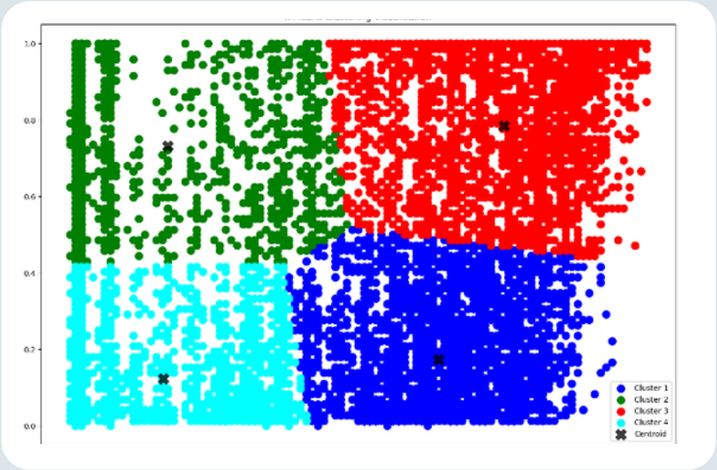
Dimensionality Reduction

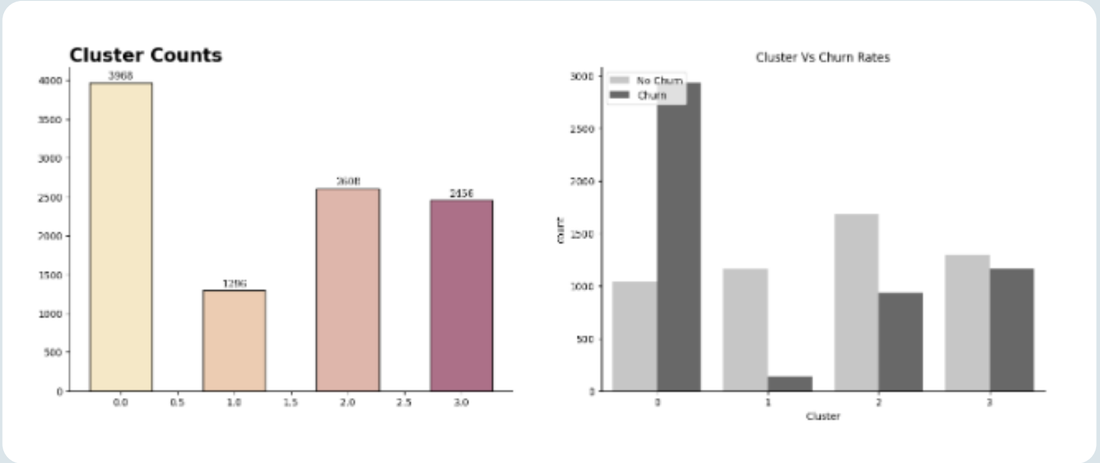
- **Purpose:** Creates a hierarchy of clusters that can either merge smaller clusters into larger ones (agglomerative) or split larger clusters into smaller ones (divisive).
- **How it Works:** Builds a dendrogram (tree-like structure) showing how data points are merged or split, with clusters being formed by cutting the tree at a desired level.
- **Key Advantage:** Does not require specifying the number of clusters in advance and provides a full hierarchy of clusters.

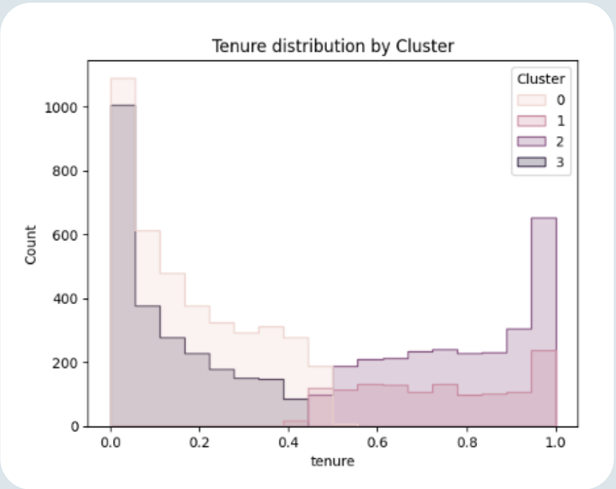
K-Means Clustering

In our analysis, we began by applying K-Means clustering to the features "MonthlyCharges" and "tenure" from the dataset. By plotting the inertia for different numbers of clusters, we identified an elbow point at around four clusters, indicating this as the optimal number. We then trained the K-Means model with four clusters and evaluated its performance using the silhouette score. The results were visualized through scatter plots, where we highlighted the clusters and their centroids. We assigned cluster labels to the dataset and analyzed the cluster counts, discovering that Clusters 0 and 1 had higher churn rates compared to Clusters 2 and 3. Further exploration revealed that Clusters 0 and 1 were associated with shorter tenures and specific characteristics such as month-to-month contracts, higher percentages of senior citizens, and a lack of online security services. After applying Principal Component Analysis (PCA) to reduce dimensionality, we repeated the K-Means clustering, once again finding four clusters as optimal. The results mirrored our earlier findings, with visualizations reinforcing the distinct characteristics across clusters. Overall, this analysis provided valuable insights into customer behavior, particularly regarding churn tendencies based on demographic and service-related features.



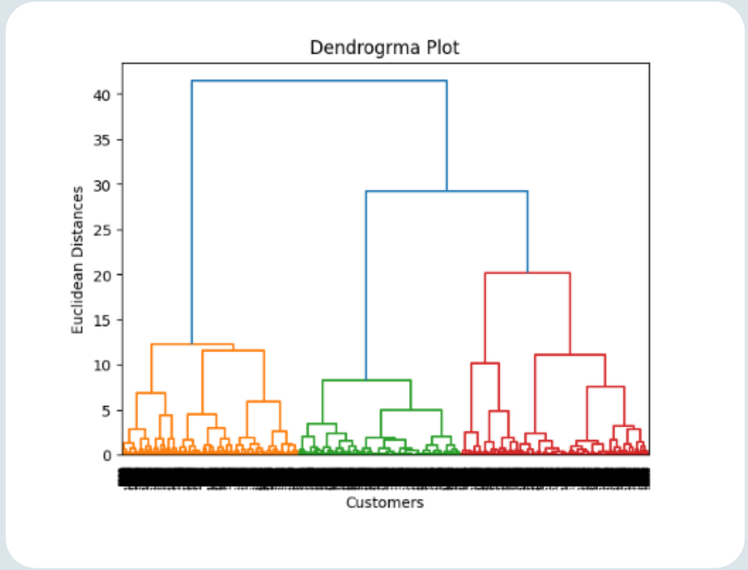


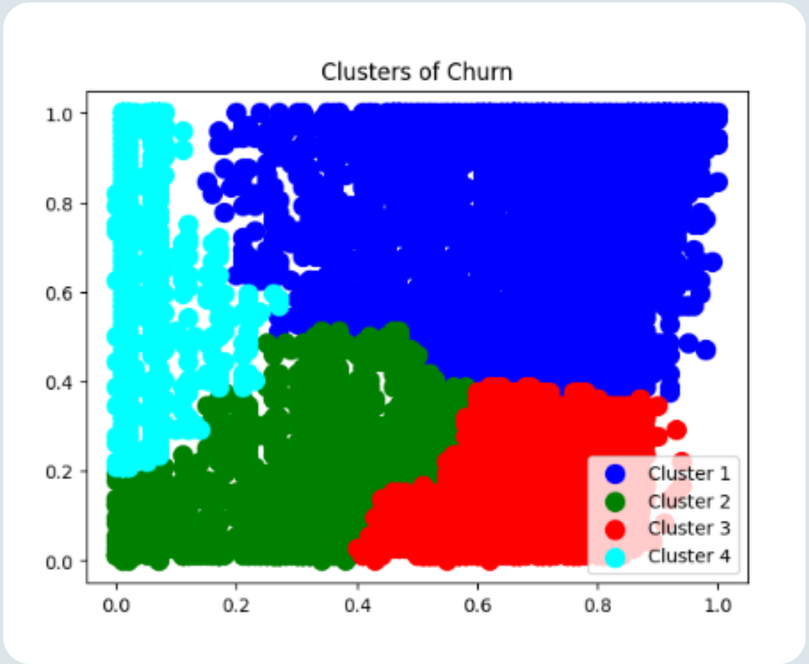


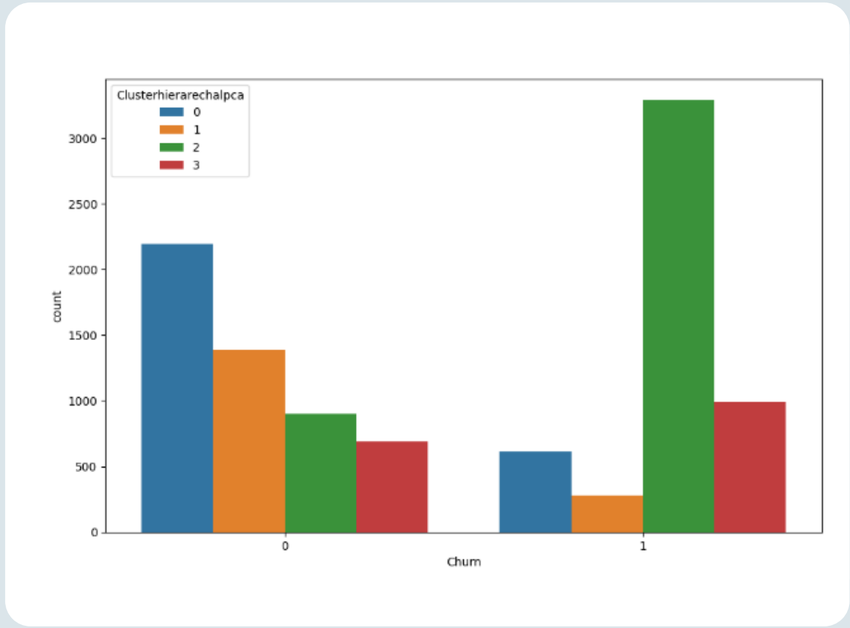


Hierarchical Clustering

In this analysis, we employed hierarchical clustering techniques to explore customer segmentation based on "MonthlyCharges" and "tenure." We began by generating a dendrogram using Ward's method, which visually represented the relationships among customers. Following this, we applied Agglomerative Clustering with four clusters, assessing its performance through the silhouette score. The results were visualized with scatter plots, effectively illustrating the distinct clusters. Each customer's cluster assignment was stored in the DataFrame for further analysis. We then examined the distribution of demographic and service-related features across the clusters using count plots, revealing insights into customer characteristics. After applying Principal Component Analysis (PCA) to reduce dimensionality, we repeated the hierarchical clustering process and visualized the results similarly. The dendrogram for the PCA-transformed data reaffirmed the appropriateness of four clusters, and the subsequent scatter plots demonstrated the clustering outcomes. This comprehensive approach provided valuable insights into customer behavior, particularly regarding churn tendencies across different demographic groups.







Business Insights

- Customer Tenure: Shorter tenures in these clusters suggest that new customers may not be finding value in the service.
- Service Utilization: The absence of online security services in higher-risk clusters indicates a potential gap in customer needs.
- Demographic Insights: A higher percentage of senior citizens in these clusters points to the need for tailored communication and services that cater to this demographic's unique requirements.
- Contract Strategies: Customers on month-to-month contracts are more likely to churn.
- Payment Preferences: The analysis indicates a preference for simpler payment methods among higher-risk segments.

Recommendations

The Telco company should:

- Give more attention to technical support
- Improve the fiber optic service
- Invest in marketing strategies targeting
- Customers with short-term contracts, trying to move them to long contracts
- Customers without online services, offering these services
- Single customers, since their churn rate is higher than that of those who have partners and dependents
- Guide customers towards simple paying methods like paper billing and credit card

Thank you