CONNECTTEL TELECOMM COMPANY

CUSTOMER CHURN PREDICTION PROJECT

PREPARED BY

TARI ATIKPA

9TH DECEMBER, 2023.

PROJECT DESCRIPTION

ConnectTel, a leading telecommunications company, is facing a significant challenge in customer churn. To address this issue, the company has initiated a project to develop a robust customer churn prediction system.

As a Data Scientist, my role in this project is to leverage advanced analytics and machine learning techniques on available customer data to accurately forecast customer churn.

The goal is to implement targeted retention initiatives, reduce customer attrition, enhance customer loyalty, and maintain a competitive edge in the dynamic telecommunications industry.

DATA VERIFICATION

- The dataset contains various features related to telecom customer information.
- Senior Citizen indicates that around 16.2% of the customers are senior citizens.
- Tenure has a range from 0 to 72 months.
- Monthly Charges range from 18.25 to 118.75.

I Performed a comprehensive check for data quality issues, including missing values and duplicates, to ensure the integrity of the dataset:

1. The result showed zero missing values across all features, indicating that the dataset is complete with no information gaps.
2. The output revealed no duplicated rows, confirming the uniqueness of each record.

PROJECT STEPS

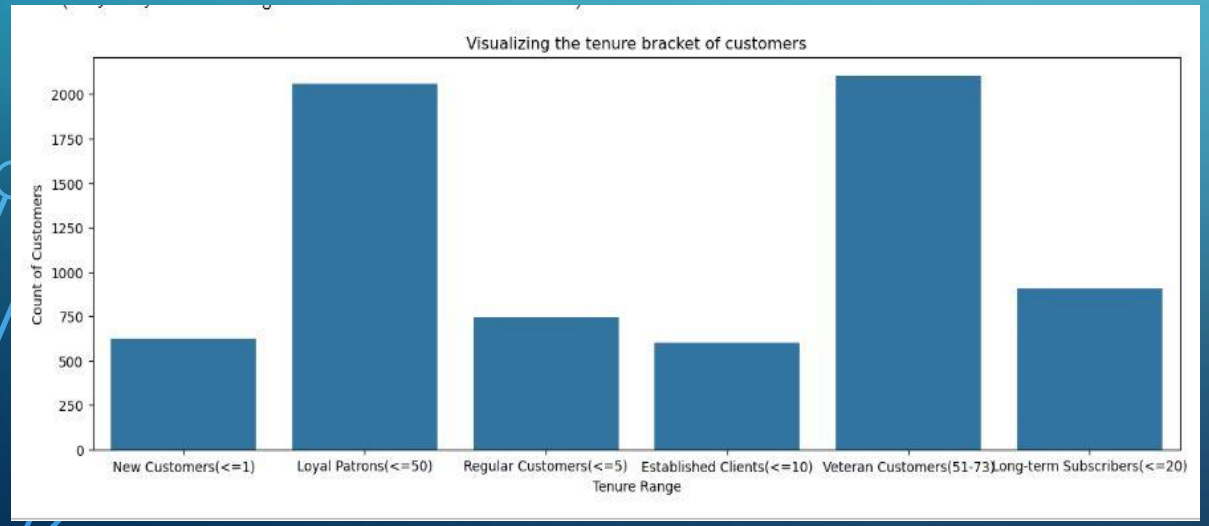
- 1. Problem Definition
- Clearly articulated the problem of customer churn and its potential impact on ConnectTel's business sustainability and growth. Defined the objective of developing a customer churn prediction system to improve customer retention strategies.
- 2. Exploratory Data Analysis (EDA) in Python
- Performed in-depth exploratory data analysis using Python. Visualized relationships between the label (churn) and key features, explored correlations, and conducted univariate, bivariate, and multivariate analysis. Identified patterns and insights to inform further steps in the project.

PROJECT STEPS

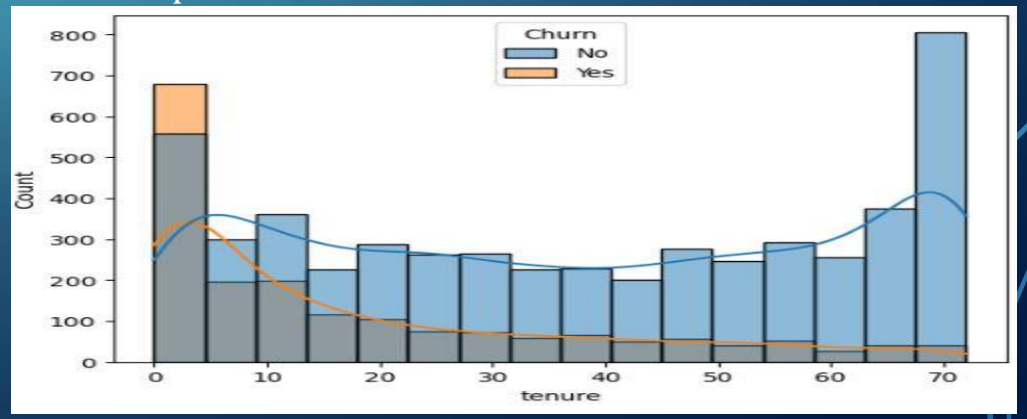
- 3. Feature Engineering
- Encoded categorical variables and created new features based on insights from EDA. Transformed the data to ensure it is suitable for machine learning model training.
- 4. Model Selection, Training, and Validation
- Selected and trained at least three supervised learning models, including Logistic Regression and Random Forest Classifier. Split the dataset into training and testing sets, trained the models, and validated their performance.
- 5. Model Evaluation
- Analyzed the results of the trained models, considering important metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Assessed the trade-offs and identified the model(s) that best address the customer churn prediction problem.

EXPLORATORY DATA ANALYSIS (EDA):

- Majority of customers are loyal patrons and veteran customers (as seen in the tenure bracket visualization).
- Senior citizens have a lower count, and there is some variability in the distribution.
- Veteran customers (51-72yearss) have the least churn, while Loyal patrons (<=50years) have the highest churn.

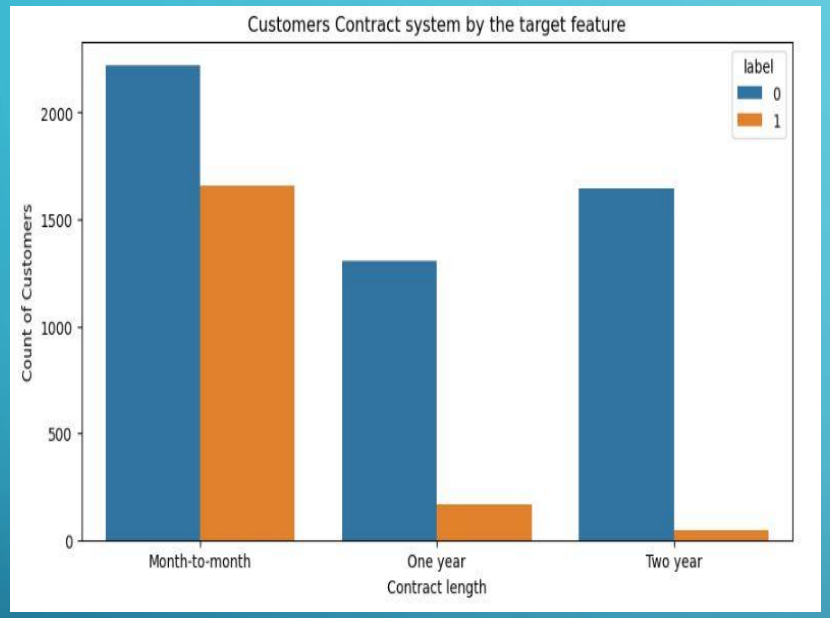


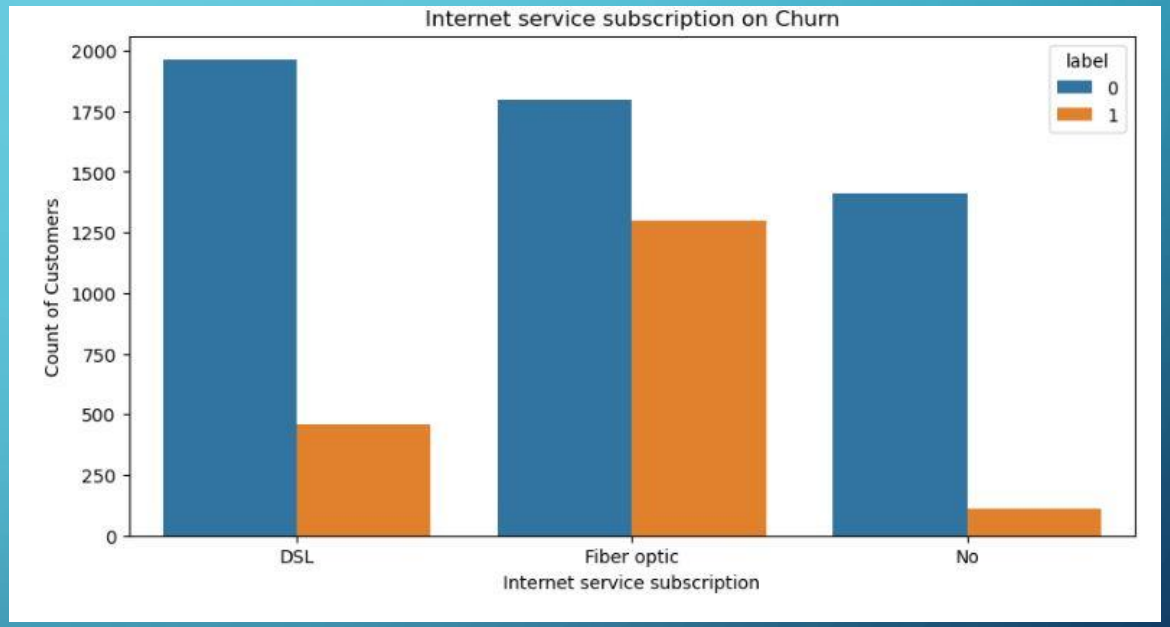
- Customers without online security services have the highest churn rate.
- Gender distribution is slightly skewed, with slightly more males.
- The distribution of tenure exhibited a right-skewed pattern, indicating that a significant portion of customers has relatively lower tenure durations. This right-skewed distribution implies that a substantial number of customers are relatively new or have shorter relationships with the telecom service.



INSIGHTS

- The impact of streaming services, contract length, internet service, paperless billing, and payment methods on churn is visualized.



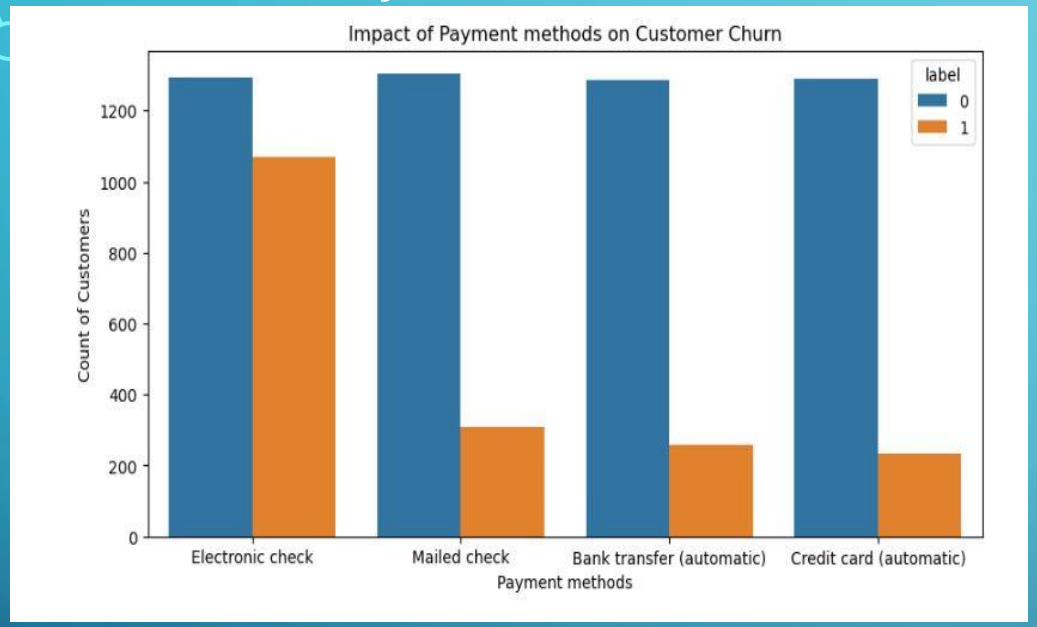


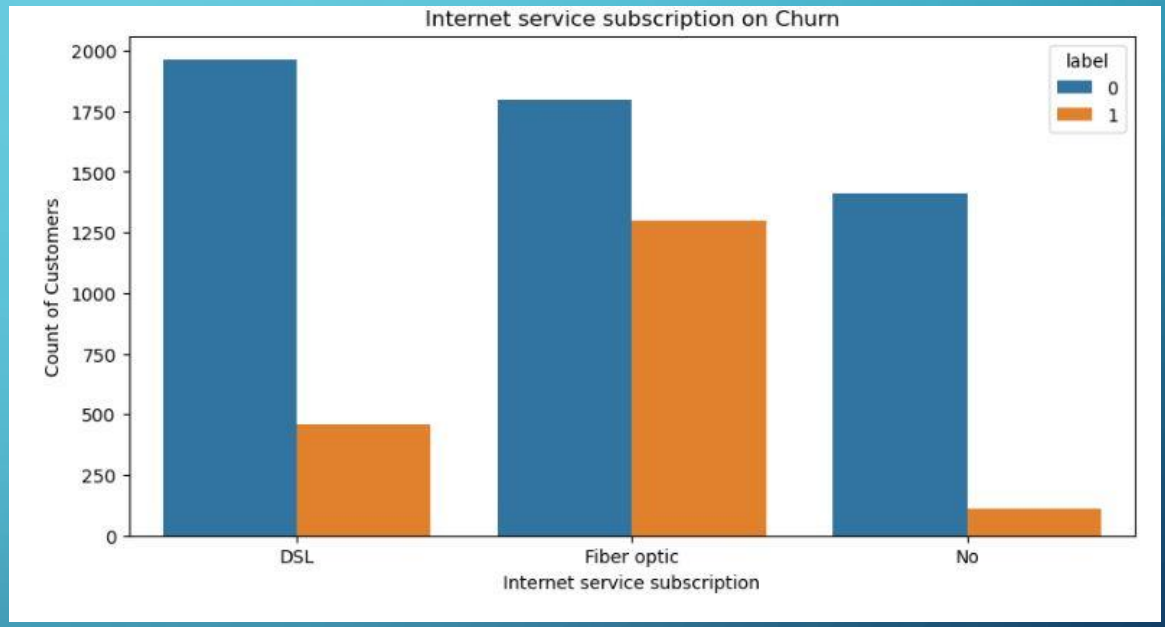
- Indicates that Customers with fibre optic service have the highest churn rate.
- Indicates that customers with no internet service have the lowest Churn rate.
- Customers with the month-to-month contract have the highest churn rate.

INSIGHTS

- The impact of streaming services, contract length, internet service, paperless billing, and payment methods on churn is visualized.

Bivariate Analysis





- Indicates that Customers who use the electronic check payment method have the highest churn rate while customers who use the credit card have the lowest Churn rate.
- Customers using fibre optic have the highest churn rate.

MODEL EVALUATION

1. Logistic Regression:

- Accuracy: The model correctly predicted approximately 79.63% of the instances.
- Precision: The model achieved a precision of approximately 64.73%, indicating that when it predicts a positive outcome (churn), it is correct about 64.73% of the time.
- Recall: The model captured approximately 50.67% of the actual positive cases.
- F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. In this case, the F1-score is approximately 56.84%.
- An AUC-ROC value of 0.7036 suggests a moderate level of discrimination ability by the model.

2. Random Forest Classifier:

- Accuracy: The model correctly predicted approximately 77.22% of the instances.

- Precision: The model achieved a precision of approximately 77.22%, indicating that when it predicts a positive outcome (churn), it is correct about 77.22% of the time.

- Recall: - The model captured approximately 77.22% of the actual positive cases.

- F1-score: The F1-score is approximately 77.22%.

- AUC-ROC: An AUC-ROC value of 0.7722 suggests a very good level of discrimination ability by the model.

The Random Forest Classifier model demonstrates strong performance across various metrics, with high accuracy, precision, recall, F1-score, and AUC-ROC. This indicates that the model is effective in making accurate predictions for both positive and negative cases.

BUSINESS RECOMMENDATIONS

- The Logistic Regression model has the highest ROC score at 70.38%. This means that the model is 70.38% accurate at distinguishing between churners and non-churners.

Overall, the results suggest that all of the models are able to distinguish between churners and non-churners with reasonable accuracy.

- The company should be more concerned with reducing false negatives. This is because false negatives can lead to lost revenue and customer loyalty. When a customer churns, the company loses the recurring revenue that the customer was generating. Additionally, churned customers are more likely to leave negative reviews and spread negative word-of-mouth, which can damage the company's reputation.
- To reduce false negatives, the company should focus on improving the recall of their churn prediction model.
- The company should identify the customers who are at risk of churning based on the model's predictions. The company can then intervene early and try to prevent these customers from churning. For example, the company could offer these customers a discount, a special promotion, or personalized customer service.

THANK YOU.