Customer churn is the rate at which customers stop doing business with a company over a given period. High churn rates can indicate customer dissatisfaction or better competition. ConnectTel, a telecom company, faces a pressing need to address customer churn which poses a significant threat to business sustainability and growth. So, predicting churn is critical for improving customer retention in ConnectTel

The objective of the analysis were to;

* Examine the relationship between churn and selected socio-economic variables.
* Develop a robust customer churn prediction system.
* Identify and visualize the most important variables that affect customer churn.
* Make recommendations to connectTel, based on findings from the analysis.

Data was obtained from the company, ConnnectTel Telecom Company and was of the type; Float64(1), Integer(2), Object(18). There was an overview of the data to check for discrepancies. Key variables in the data were gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges and Churn. User ID was removed from the analysis as it wasn’t a variable per se.

Next I did was to carry out exploratory data analysis on selected variables. Univariate analyses on monthly charges reveals box plot with data of spread 20-120 being is skewed. Histogram shows frequency of 1600 when the charges was just between 20 and 30. Further, the graph showed a normal distribution for monthly charges that ranged between 30 and 120 with highest frequency being 1000

Further univariate analysis was carried out with Payment method. The histogram reveals that the electronic payment has the most frequency of more than 2000 compared to other forms of payment. Boxplot for Tenure was skewed with median at 30. It was also observed that Tenure does not follow normal distribution in the histogram, giving a pointer to it as a variable to watch out for.

The Bivariate Analysis revealed that more than 1000 customers that use electronic check payment method had churned. More than 1250 customers that use fiber optic internet services had chuned.

Further analysis was carried out using correlation. The result revealed that there is a moderately weak and negative linear relationship between churn and contract with coefficient of -0.40.

In addition to the above analysis, three machine learning models where trained and tested

**Model Performance**

* **Logit**.
* Strengths: High accuracy (81.48%) and a good AUC-ROC score (0.7351), indicating that it performs well overall in predicting customer churn. The precision (68.06%) is also relatively high, meaning that when the model predicts a customer will churn, it is often correct.
* Weaknesses: The recall (56.57%) is somewhat lower, indicating that the model does not catch all churners. This suggests that there might be room for improving the model to better capture those customers who will actually churn.

**Random Forest.**

* Strengths**:** Reasonably high accuracy (79.13%) and moderate precision (64.79%), indicating that it performs well overall and is fairly reliable when it predicts churn.
* Weaknesses**:** The recall (46.38%) is relatively low, indicating that the model does not catch a significant portion of actual churners. The F1-score (0.5406) and AUC-ROC (0.6865) suggest that there is considerable room for improvement in the model's performance

**SGDClassifier**

* Strengths**:** The model has high accuracy (81.48%) and good precision (68.06%), suggesting it performs well overall and is reliable when it predicts churn. The AUC-ROC score (73.51%) indicates that the model is effective at distinguishing between churners and non-churners.
* Weaknesses: The recall (56.57%) is relatively low, indicating that the model misses a significant portion of actual churners. The F1-score (0.6179) reflects this imbalance between precision and recall.

The other thing I carried out was feature importance using random forest classifier. The choice of random forest was due to the fact that its confusion matrix gave the lowest value of the three models. A low false negative was key to this analysis.

**Insights, Recommendations and conclusion**

Since there is a weak negative correlation between churn and contracts with value of -0.04. It is recommended that the company would strengthen its contract terms.

The confusion matrix from the Random Forest reveals a lower value for FALSE NEGATIVE compared to the other two models. Though this will pose a lesser risk to the company compared to other two models, the model still fails to identify a good number of customers who are actually at risk of churning.

This is disturbing because it means that the company may not be taking proactive measures to retain these customers, potentially leading to revenue loss and reduced customer satisfaction.

However, due to the insights gotten from the feature importance, targeted retention strategies and interventions for customers identified as high-risk churners would be developed