PHASE 3 ~ PROJECT

FOCUS:

Machine Learning Fundamentals

PROBLEM STATEMENT

- Develop a predictive model to identify customers at risk of churning (leaving the service) for SyriaTel, a telecommunications company. This binary classification model will help proactively detect high-risk customers, enabling targeted retention strategies to reduce customer churn and improve business performance.

BUSINESS OBJECTIVES

- Churn Prediction: Identify the factors that are most likely to lead to customer churn.
- Customer Retention: Develop a model that can accurately predict which customers are at risk of churning. SyriaTel can take proactive steps to retain customers who are at risk of churning.
- Cost Reduction: Predicting churn allows SyriaTel to allocate resources effectively, targeting the customers who require intervention before they leave. This reduces the costs associated with acquiring new customers to replace those lost.
- Revenue Retention: Ultimately, reducing churn will contribute to higher customer lifetime value (CLV) and revenue retention, ensuring the long-term profitability and sustainability of SyriaTel.

DATA UNDERSTANDING

- The dataset has 3333 rows and 21 columns
- The datatypes of the dataset include: bool(1), Float(64), Int64(8), Object(4)
- The data has two columns: the numerical and categorical columns

* Numerical Columns: account length , area code , number vmail messages, total day minutes, total day calls, total day charge, total eve minutes, total eve calls, total eve charge, total night minutes, total night calls, total night charge, total intl minutes, total intl calls, total intl charge, customer service calls

* Categorical Columns: state, phone number, international plan, voice mail plan

DATA CLEANING: The Data contained no missing values or duplicates.

EDA(EXPLORATORY DATA ANALYSIS)

- Univariate Analysis:
- Bivariate Analysis: The relationship between two variables : Customer Service Calls and Churn , total intl charge and International, total night charge and night, total eve charge and evening,
- Univariate Analysis: To describe a type of data which consists of observations on only a single characteristic or attribute: Churn Distribution, Distribution of Area Code Feature, Numerical Features and Categorical Features

MODELING

- The algorithms used include:

- Logistic Regression

- Decision Tree

- Random Forest

- We will also be using the `ROC_AUC `metric to evaluate the performance of our models
- To deal with class imbalance, we will be using `SMOTE` to generate synthetic examples of the minority class in our dataset
- Train and Split the Dataset
- Scaling the data: Used the MinMaxScaler since the effect is it scales to [0,1] and handles no outlier, used when the feature has a varying scale.
- Synthetic Minority Oversampling Technique ("SMOTE") is an oversampling technique where synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

LOGISTIC REGRESSION

- Its used as a baseline model which has a recall score of 51%, considered the absolute minimum to provide meaningful insights.
- The model correctly approximately identifies 51% of the actual positive instances.

	Precision	Recall	F1-Score	Support
0	0.90	0.88	0.89	664
1	0.44	0.51	0.47	129
Accuracy Macro Avg Weighed Avg	 0.67 0.83	 0.69 0.82	0.82 0.68 0.82	793 793 793

DECISION TREE CLASSIFIER

- It is a supervised machine learning algorithm that can be used to classify data. Decision trees work by splitting the data into smaller and smaller subsets until each subset contains only data of a single class. The decision tree then predicts the class of a new data point by following the path down the tree that corresponds to the values of its features
- The Decision Tre classifier model has a recall score of 0.75. This means that the model can identify around 75% of the actual positive instances correctly.

	Precision	Recall	F1-Score	Support
0	0.95	0.91	0.93	664
1	0.62	0.75	0.68	129
Accuracy Macro Avg Weighed Avg	 0.79 0.90	 0.83 0.89	0.89 0.81 0.89	793 793 793

RANDOM FOREST CLASSIFIER

- It is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting a class prediction or regression value by averaging the predictions of the individual trees.
- The random forest classifier model has a recall score of 0.64. This means that the model can identify around 64% of the actual positive instances correctly.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

	Precision	Recall	F1-Score	Support
0	0.93	0.97	0.95	664
1	0.80	0.64	0.71	129
Accuracy Micro Avg Weighed Avg	 0.87 0.91	 0.80 0.92	0.92 0.83 0.91	793 793 793

MODEL EVALUATION

- **Models Comparison – Recall Score**
- Evaluate models based on recall score and ROC_AUC. After, we will the best two models to tune them for better performance .
- The recall score is a measure of how many of the positive instances the model correctly identifies. A higher recall score indicates that the model is better at identifying positive instances.
- The DecisionForestClassifier has the highest recall score, followed by RandomForestClassifier and, LogisticRegression.The LogisticRegression has the lowest recall score of 0.511628
- **Models Comparison – ROC Curve**
- The ROC curve analysis shows that the RandomForestClassifier has the best performance, followed by the DecisionTreeClassifier, and LogisticRegression. The RandomForestClassifier has the highest AUC score of 0.885, while the LogisticRegression has the lowest AUC score of 0.753.
- A higher AUC score indicates that the classifier is better at distinguishing between positive and negative instances.

MODEL TUNING

- Based on the evaluation of the models, it is observed that the RandomForest classifier and the Decision Tree classifier have shown promising performance.
- To further improve their performance, model tuning can be performed using GridSearch.
- The tuned Decision Tree model performs well, the model has a recall score of 0.75, which means model is able to capture 75% of the actual churned customers.
- The tuned Random Forest model has a recall score of o.67, which means model is able to capture 67% of the actual churned customers.

RECOMMENDATIONS

- **Target High-Churn Areas with Incentives:**

~Offer discounts or promotional deals to customers in area codes 415 and 510, which have shown higher churn rates.

- **Enhance Customer Service Quality:**

~Improve training programs for customer service representatives to ensure prompt, effective issue resolution.

- **Reevaluate Pricing Structures:**

~ Review the pricing for day, evening, night, and international calls.

~Adjust plans or introduce discounted packages to better align with customer needs and address cost concerns that may contribute to churn.

- **Focus on High-Churn States:**

~ Implement targeted retention strategies in states with elevated churn rates, such as Texas, New Jersey, Maryland, Miami, and New York.

- **Promote Voicemail Plan Adoption:**

~Improve the value proposition of voicemail services by highlighting their convenience and benefits.

NEXT STEPS

- Implementation of the various recommendations depending on the resources and based on the priority in the business.

THANK YOU