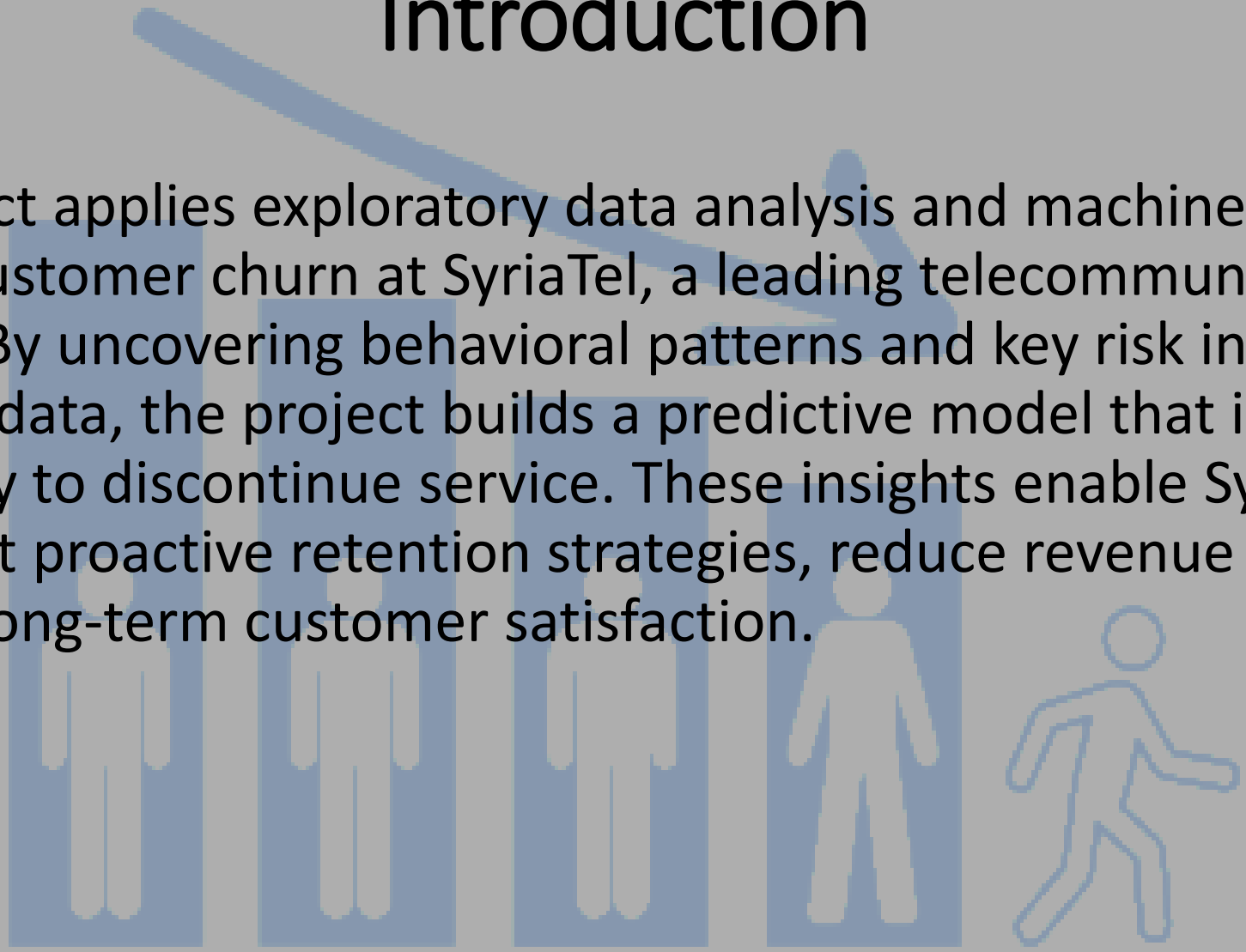
Churn Prediction in Telecommunications Using Logistic Regression and Decision Trees

PHASE 3 Project

Author: Danton Kipngeno

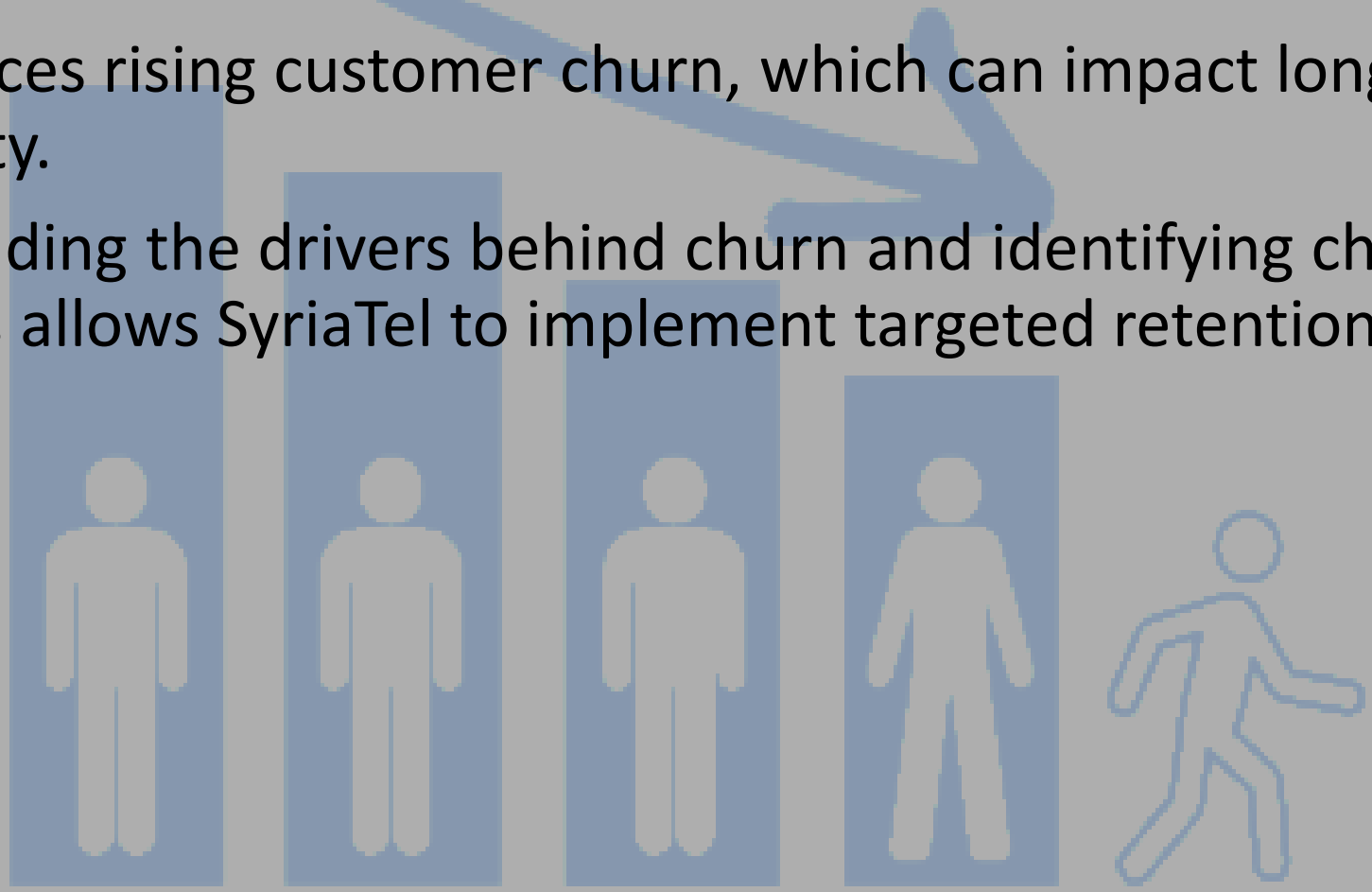
Introduction

- This project applies exploratory data analysis and machine learning to address customer churn at SyriaTel, a leading telecommunications provider. By uncovering behavioral patterns and key risk indicators in customer data, the project builds a predictive model that identifies users likely to discontinue service. These insights enable SyriaTel to implement proactive retention strategies, reduce revenue loss, and enhance long-term customer satisfaction.



Business problem

- SyriaTel faces rising customer churn, which can impact long-term profitability.
- Understanding the drivers behind churn and identifying churn-prone customers allows SyriaTel to implement targeted retention strategies.



Key Business questions

- What customer attributes are most predictive of churn?
- Are there usage behaviors that indicate higher churn risk?
- Can we build a model that accurately identifies customers who are likely to churn?
- How can SyriaTel use this model to intervene and retain at-risk customers?

Data Overview

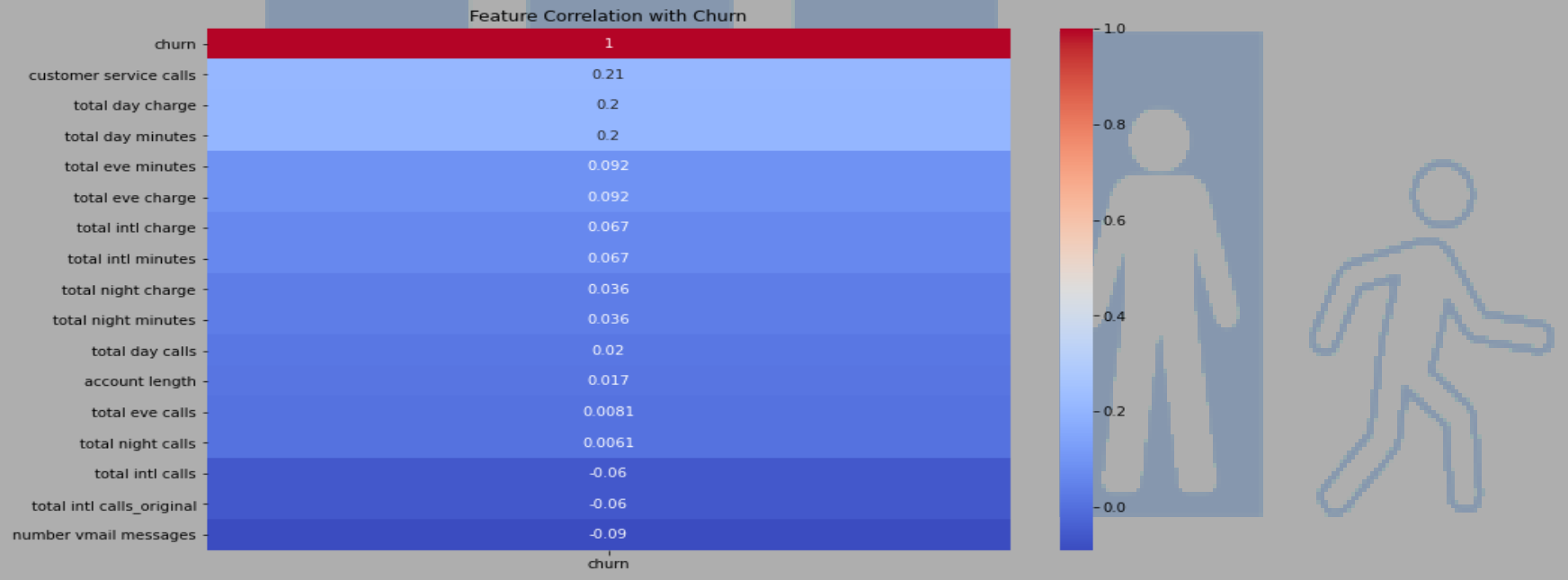
- The SyriaTel Customer Churn dataset contains 3,333 records and 20 features, with Churn as the binary target variable.
- After removing non-informative columns (Phone Number, State, Area Code), the data was cleaned and found to have no missing values. Key categorical features (e.g., International Plan, Voice Mail Plan) were encoded, and numerical features were standardized.
- Outliers were identified in Customer Service Calls, Total International Calls, and Total International Charge.
- The dataset is imbalanced, with only 14.5% churned cases, addressed using SMOTE and class weighting. An 80/20 train-test split was used for model evaluation.

Feature Selection & Engineering

➢Identifying the Most Predictive Features:

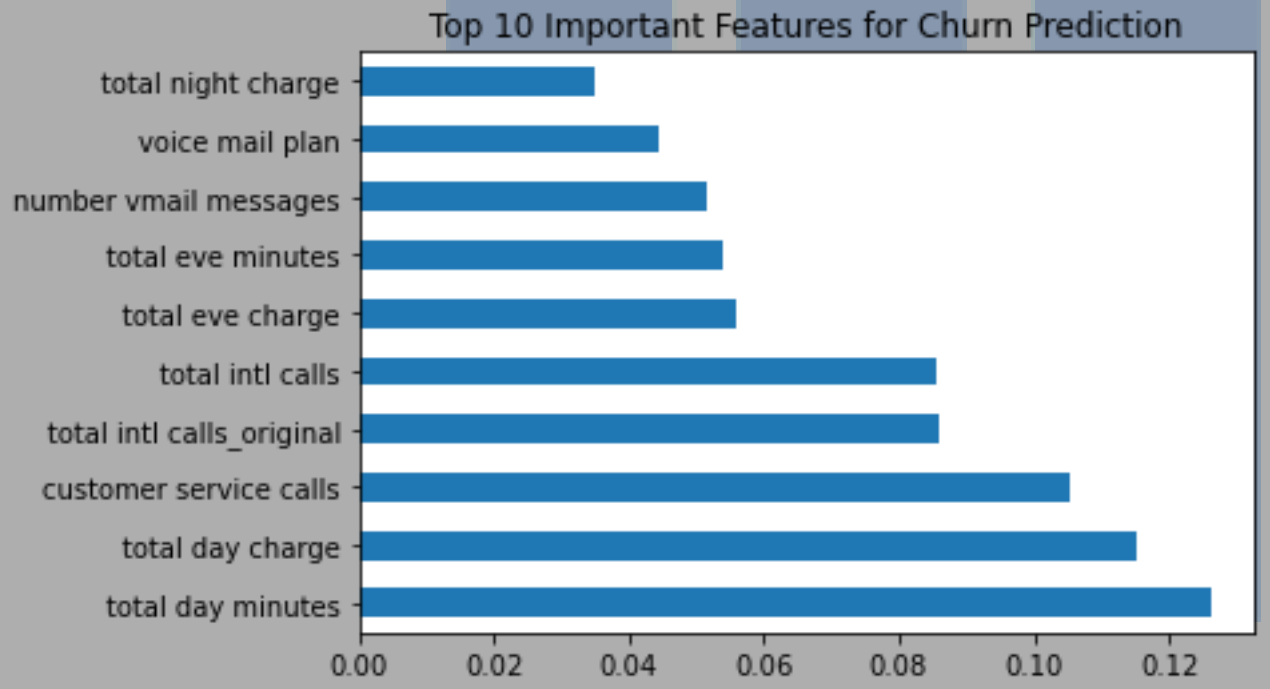
a. Correlation Analysis

- Numeric features were selected for analysis using select_dtypes.
- Pearson correlation was calculated to examine linear relationships with the churn variable. Heatmap plotted to visualize correlation strengths.



Feature Importance using Tree-based Model

- Random Forest model was trained to identify feature importance.
- Feature importance scores reflect how much each variable reduces impurity in tree splits.
- Top Features Identified include: Total day minutes, Total day charge, Customer service calls



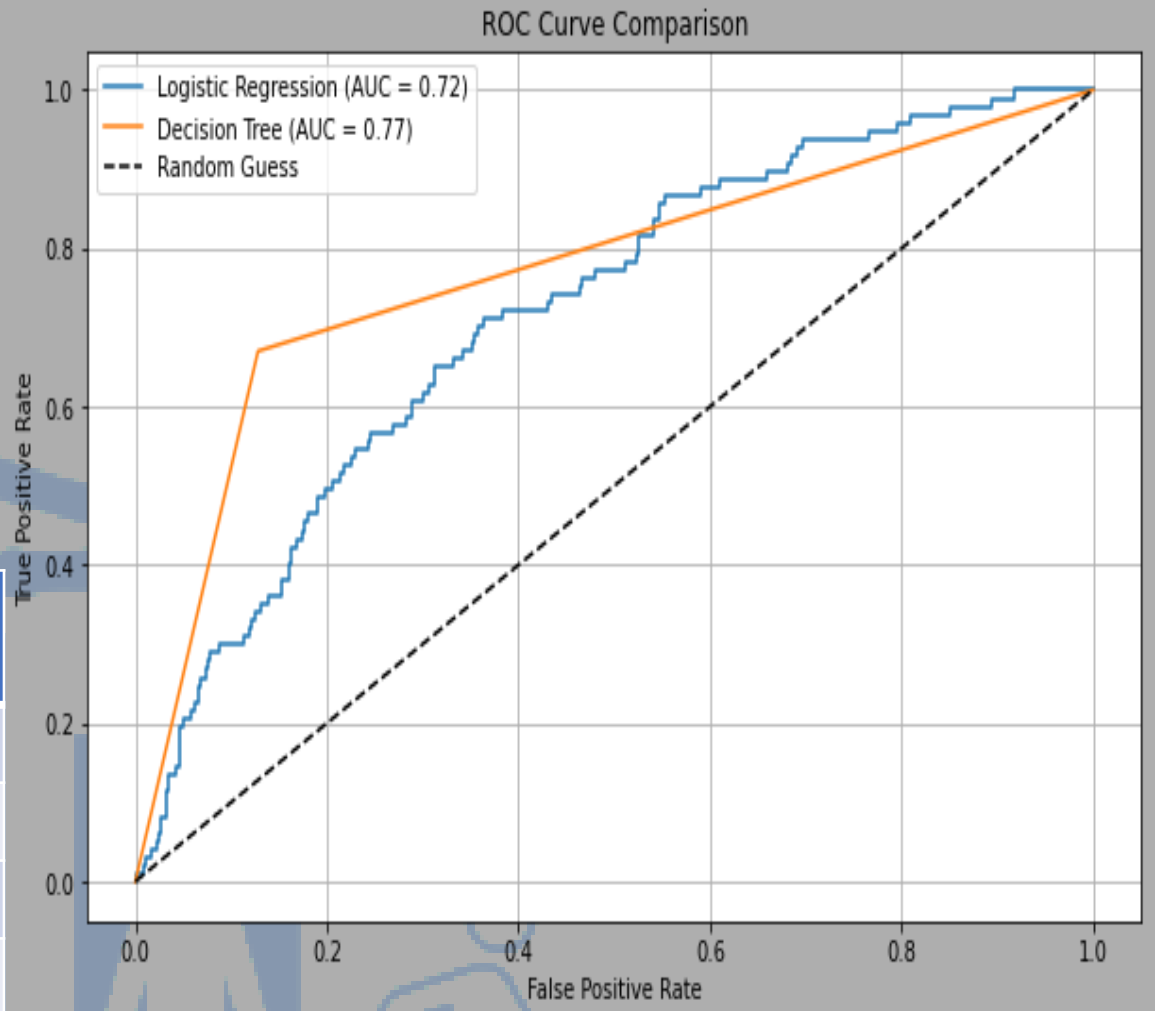
Modelling approach

Models Used

- Logistic Regression (baseline)
- Decision Tree Classifier (tuned)

Model Performance Evaluation

Metric	Logistic Regression	Decision Tree
Accuracy	66%	84%
Recall (Churn)	0.67	0.67
Precision (Churn)	0.25	0.47
AUC Score	0.72	0.77

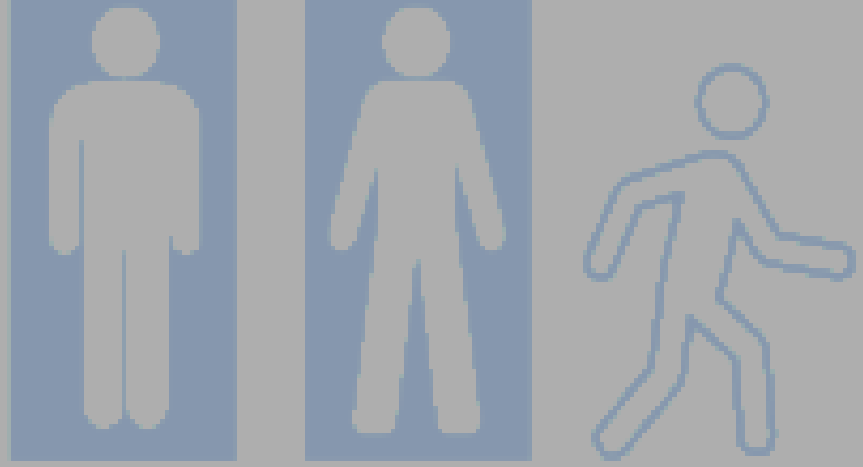


- Both models detect churners equally.
- Decision Tree is more precise and accurate.
- AUC curve confirms superior class distinction by Decision Tree.

Overfitting Assessment

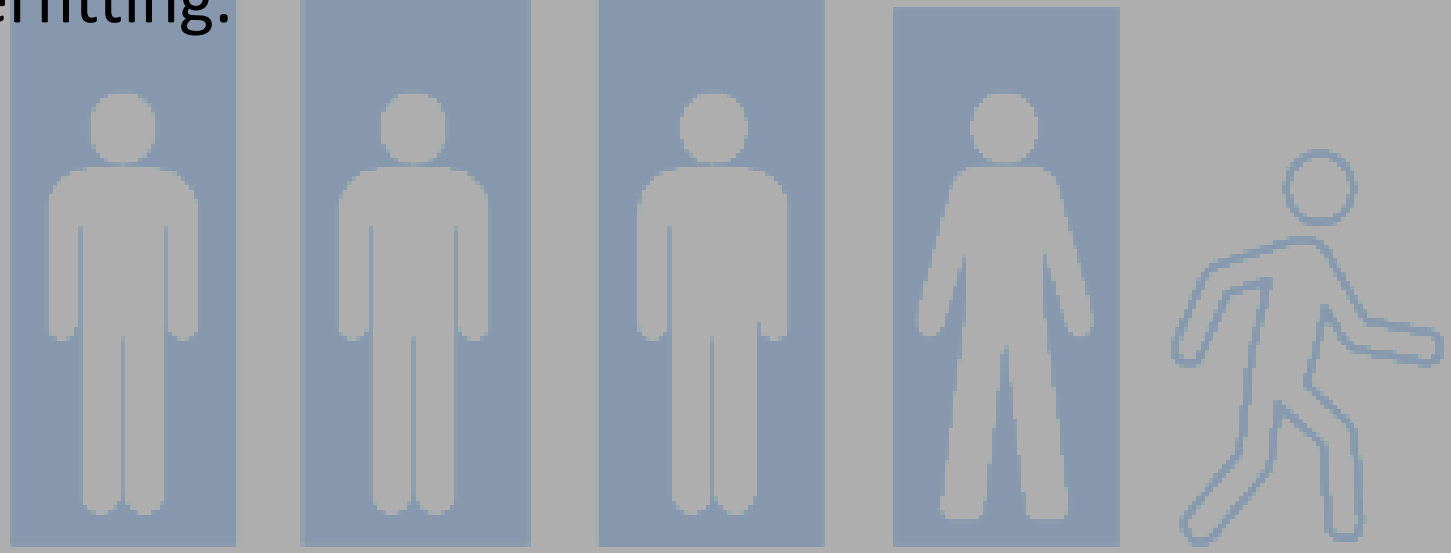
- **Training Set Performance** (Decision Tree): Accuracy of **100%** and AUC of **1.0**
- **Test Set Performance**: Accuracy of **90%** and AUC: **0.87**
- This showed that Overfitting was Detected since we have perfect training performance but a notable drop on test set.
- Model memorized training data; needs tuning to generalize better.

```
Training Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2280
           1       1.00      1.00      1.00      2280

    accuracy                           1.00      4560
   macro avg       1.00      1.00      1.00      4560
weighted avg       1.00      1.00      1.00      4560

Training AUC: 0.9999999999999999
Test Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       570
           1       0.67      0.66      0.66        97

    accuracy                           0.90       667
   macro avg       0.80      0.80      0.80       667
weighted avg       0.90      0.90      0.90       667

Test AUC: 0.8720926026406223
```



Hyperparameter Tuning

- GridSearchCV used to find optimal tree parameters: max_depth, min_samples_split, min_samples_leaf, max_leaf_nodes.
- Used 5-fold cross-validation for robustness.
- After tuning we got Best AUC Score of 0.936 (Cross-validated)
- This shows that tuned model shows excellent class separation and avoids overfitting.



Evaluation

- We can observe an Accuracy of 0.88 which is Strong overall performance.
- AUC Score of 0.83 which shows Good ability to distinguish between churners and non-churners.
- Class 0 (No Churn) has excellent performance (high precision & recall).
- Class 1 (Churn) has Performance improved significantly after tuning recall is now 0.72, meaning the model captures 72% of actual churners (up from 66%).

Test Classification Report:

	precision	recall	f1-score	support
0	0.95	0.90	0.93	570
1	0.56	0.72	0.63	97
accuracy			0.88	667
macro avg	0.75	0.81	0.78	667
weighted avg	0.89	0.88	0.88	667

Test AUC Score: 0.8299963827093507

Conclusion

- Feature importance analysis revealed that "total day minutes," "total day charge," "number of customer service calls," and international plan subscription were among the most predictive features of churn.
- The analysis showed that some customers characteristics were likely to contribute to churn. People who use their mobile phones during the day for many minutes and charges, often calling the customer service centre or using an international plan are likely to cancel their service.
- We were even able to build a machine learning model that we could use to predict customers most likely to churn. To summarize, there is a high accuracy after re tuning with Decision Tree model particularly in identifying the churners with the recall rate of 72%.
- Model can be incorporated in the customer management system to alert the firm when customers are considering to leave the firm

Recommendation

- Deploy the model in SyriaTel’s operational systems to flag high-risk customers monthly or weekly.
- Train support teams to handle frequent callers with care, as they are more likely to churn.
- Direct retention incentives toward customers with high day-time usage or those on international plans
- Continuous Monitoring: Regularly retrain the model with updated data to maintain performance as customer behavior evolves.