

Customer Churn Analysis

Harry Atulah



TABLE OF CONTENTS

01

OVERVIEW

04

MODELLING

02

**BUSINESS
UNDERSTANDING**

05

EVALUATION

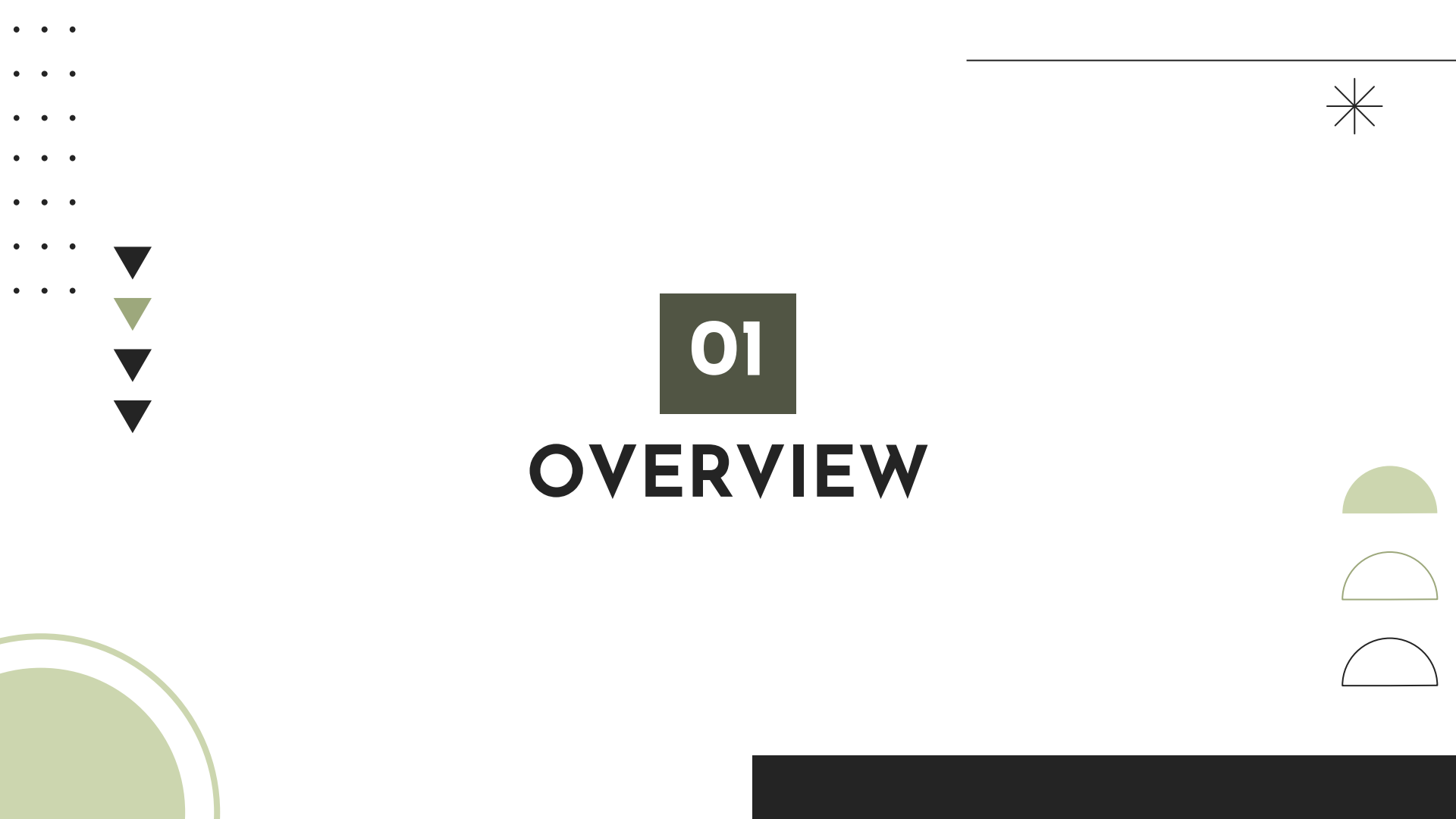
03

DATA UNDERSTANDING

06

RECOMMENDATION




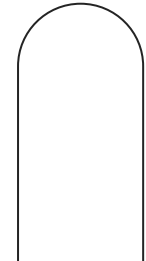


01

OVERVIEW



INTRODUCTION

- The objective of this project is to develop a predictive model to accurately identify customers at risk of churning, using a combination of data preprocessing, feature engineering, and oversampling techniques like SMOTE to address class imbalance.
 - By employing a Random Forest classifier, we aim to uncover key factors influencing churn and generate actionable insights that can help the business implement targeted retention strategies.
 - The process involves rigorous model evaluation to ensure robustness and reliability, ultimately supporting data-driven decision-making to reduce customer attrition and enhance long-term customer loyalty.
 - Methodology: Data preprocessing, feature engineering, oversampling with SMOTE, and model evaluation.
- 
- 




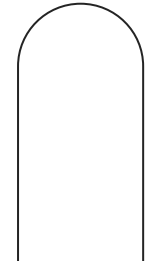
02

BUSINESS UNDERSTANDING





PROBLEM STATEMENT

- Customer churn occurs when a customer stops using the company's products or services.
 - The business problem is to develop a classifier that can accurately predict whether a customer is likely to churn or not based on various features such as usage patterns, account information, and customer demographics.
 - By identifying at-risk customers early, SyriaTel can implement targeted retention strategies to reduce churn and improve customer satisfaction and loyalty.
- 
- 



Objective

Our goal is to develop a robust predictive model that accurately identifies customers at risk of churning, utilizing data preprocessing, feature engineering, and advanced machine learning techniques performance.



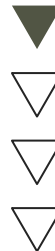
Stakeholders

- SyriaTel
- Executive
- Marketing team
- Customers
- Data Analysts



Business Impact

- Revenue loss
- Increased cost
- Market share decline
- Reputation damage





03

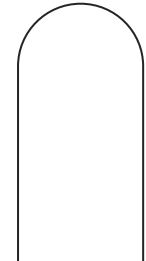

DATA UNDERSTANDING






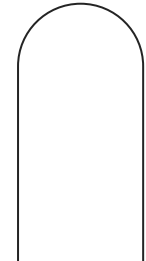
Dataset

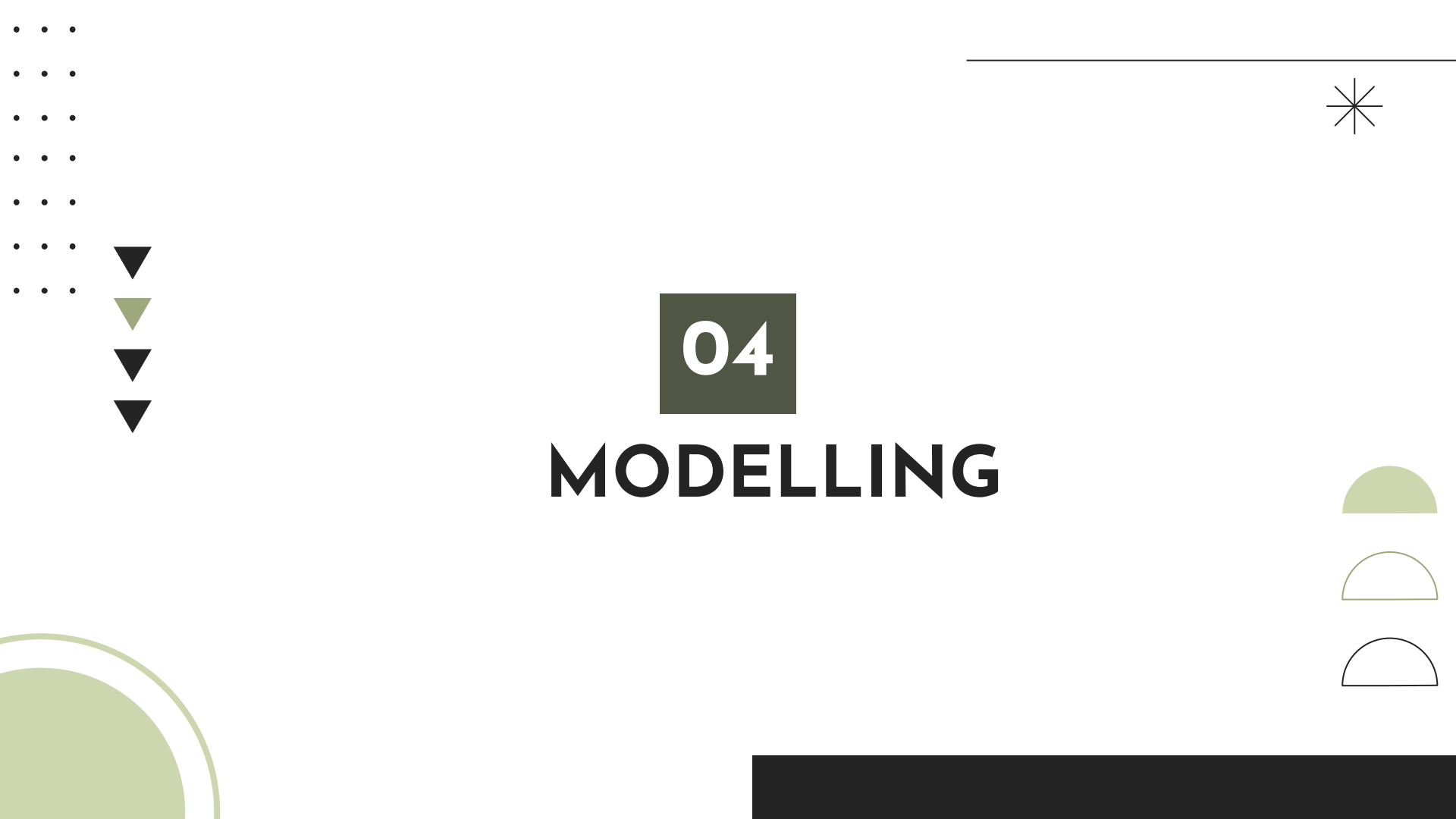
The dataset used in this analysis comprises various attributes related to telecom customers, including account length, area code, phone number, international plan subscription, voicemail plan subscription, number of voicemail messages, total minutes of usage during the day, number of calls made during the day, charges incurred during the day, and similar metrics for evening and night usage. Additionally, it includes features such as total international minutes, international call counts, and associated charges, along with the number of customer service calls made. The target variable, "churn," indicates whether a customer has churned or not. This dataset allows us to explore patterns and factors contributing to customer churn in the telecom industry..





Data Analysis

- We first explored the dataset to understand its structure and contents. This involved examining descriptive statistics, checking for missing values, and identifying any outliers.
 - Next, we conducted visualizations to gain insights into the distribution of key features, correlations between variables, and the proportion of churned customers.
 - We also performed feature engineering to create new variables or transform existing ones to enhance the predictive power of our models. Additionally, we conducted statistical tests or exploratory data analysis to identify significant factors associated with churn.
 - The data analysis phase aimed to uncover patterns and relationships within the data that could help us better understand customer churn behavior.
- 
- 


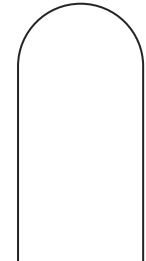


04

MODELLING


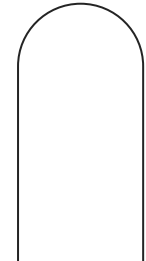


Models Used

- Logistic Regression was the first model that was used. We trained the Logistic Regression model with a random state of 42, made predictions, and printed out the results. The accuracy of the Logistic Regression model on the test set was printed, followed by the classification report and confusion matrix.
 - Next we trained a Decision Tree Classifier with a random state of 42, made predictions, and printed out the evaluation results. The accuracy of the Decision Tree model on the test set was printed, followed by the classification report and confusion matrix.
 - Finally, we trained a Random Forest Classifier and printed out the outputs.
- 
- 



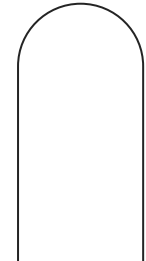

Applying SMOTE

- To address the class imbalance issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.
 - SMOTE generates synthetic samples for the minority class, balancing the distribution of classes in the dataset.
- 
- 



Hyperparameter Tuning

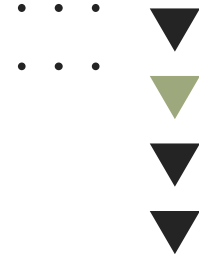
GridSearchCV was employed to find the optimal hyperparameters for the Random Forest Classifier. This process systematically searches through a grid of hyperparameters, using cross-validation to determine the best combination that maximizes model performance.






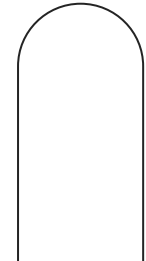
05

EVALUATION




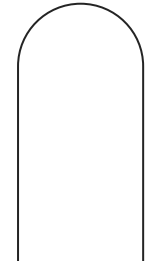


Metrics

- The evaluation of the model's performance included various metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics provide insights into different aspects of the model's predictive capabilities.
 - To compare the three models, Logistic Regression, Decision Tree, and Random Forest, we first evaluated their performance metrics on the test set. Logistic Regression achieved an accuracy of 85.5%, demonstrating good precision (87%) for the "False" class but lower recall (19%) and F1-score (28%) for the "True" class.
 - Decision Tree outperformed with an accuracy of 94.3%, showing balanced precision (95%) and recall (98%) for the "False" class and acceptable precision (88%) but lower recall (72%) for the "True" class.
 - Random Forest attained an accuracy of 89.8%, with a strong precision (89%) and perfect recall (100%) for the "False" class, but lower recall (34%) and F1-score (50%) for the "True" class.
 - The confusion matrices also highlight the model performances
- 
- 



Cont:

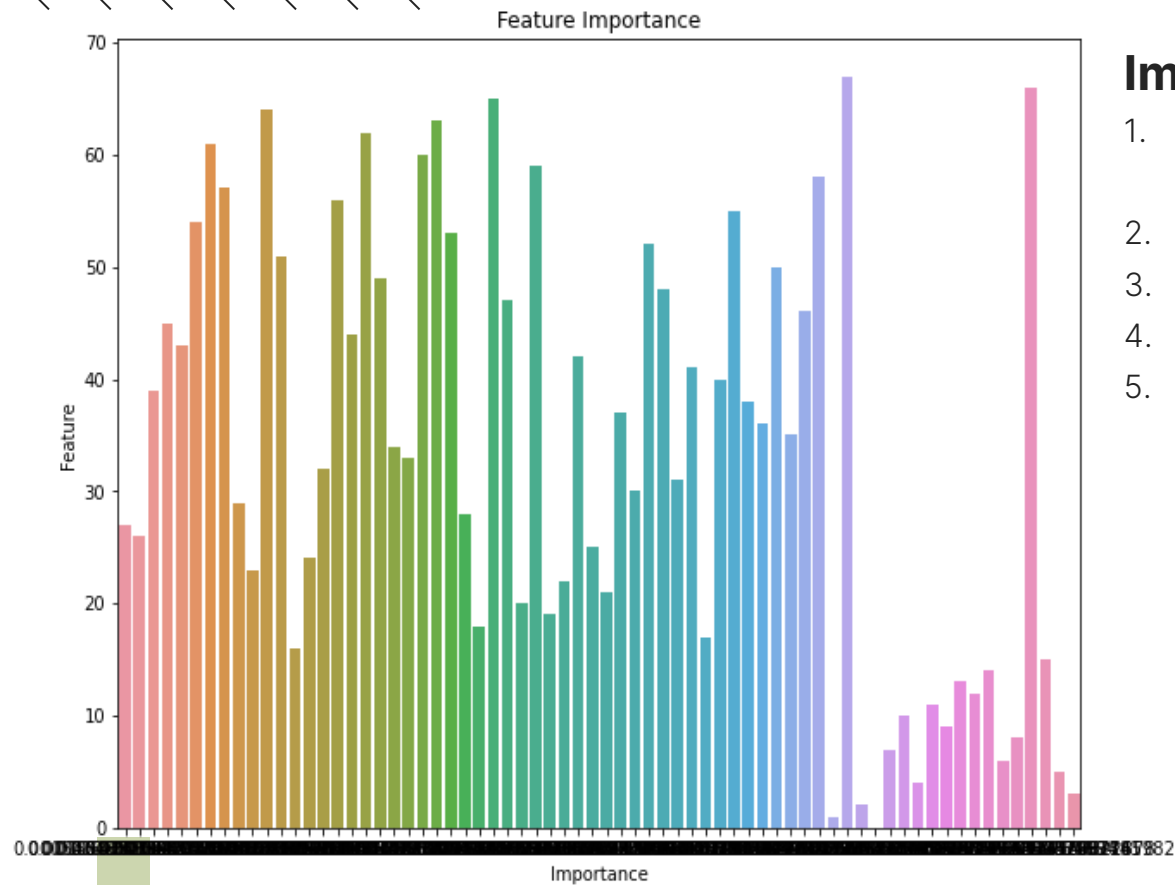
- We improved the random forest model using hyper parameter tuning where it showed a significant increase in its accuracy to predict customer churn.
 - We got an accuracy of 94.3% and a best score of 95.4% which improved our model's ability to predict the probability of a customer to stop using a company's product due to various reasons.
- 
- 



06

RECOMMENDATION




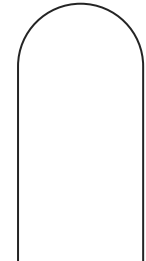


1. Total Day Minutes and Total Day Charge
2. Customer Service Calls
3. International plan
4. Total Evening Charge
5. State-Specific Insights



Recommendations

Based on the important features, these are the recommendations we gave Syriatel:

1. Syriatel should monitor high-usage customers for potential churn triggers and offer special plans or discounts to heavy users to encourage retention.
 2. Improve customer service quality and responsiveness. Track frequent callers and ensure their issues are resolved promptly and satisfactorily.
 3. Introduce attractive evening and off-peak hour plans or discounts to cater to users who predominantly use the service during these times.
 4. Analyze and address state-specific customer behaviors and issues. Customize marketing and retention strategies to cater to regional preferences and needs.
 5. Promote competitive international plans and ensure customers are aware of these options. Offer personalized discounts for frequent international callers.
- 
- 



Contact info

Name: Harry Atulah

Linkedinprofile:

<https://www.linkedin.com/in/harry-atulah-14668819b/>

