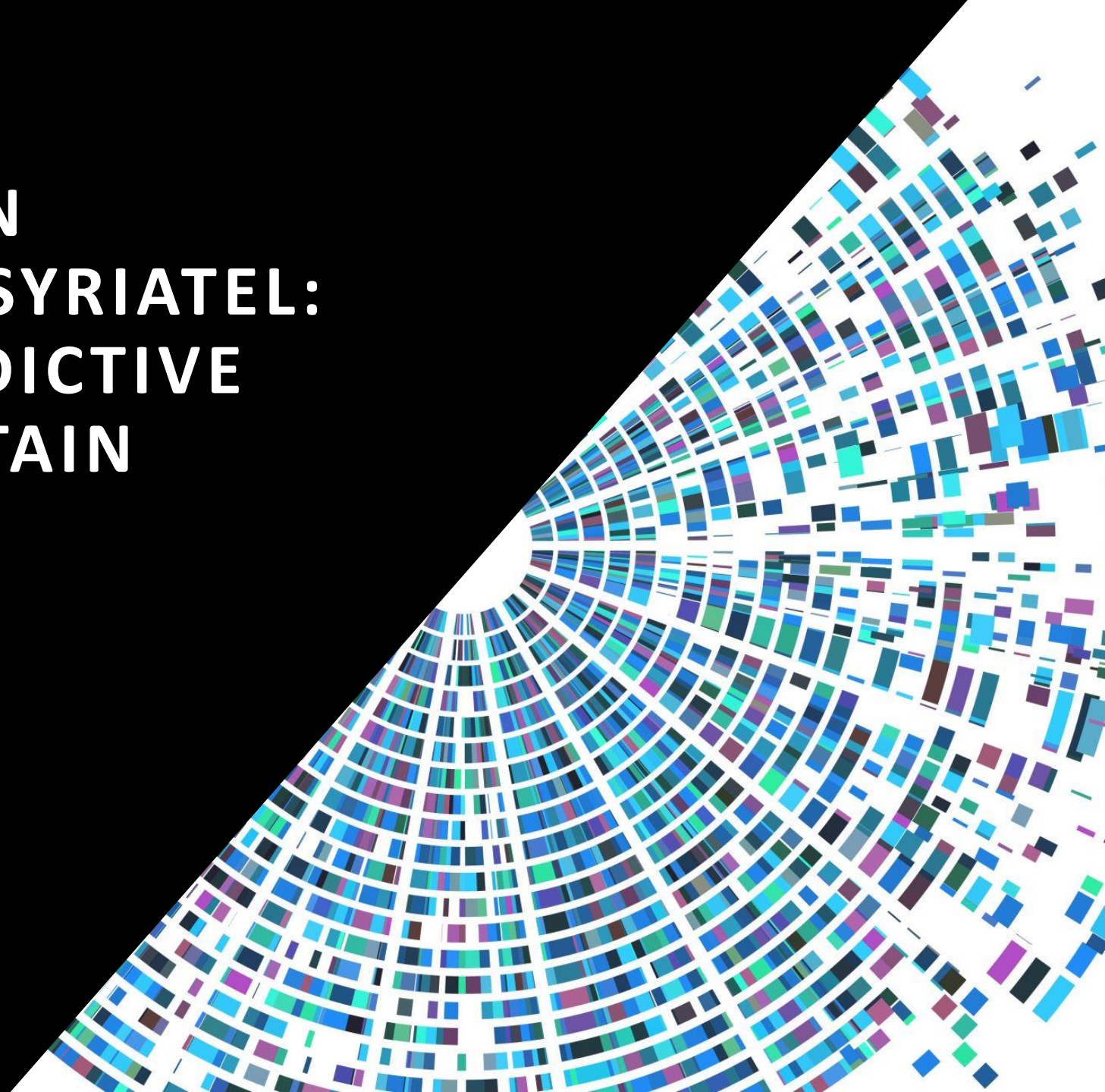
CUSTOMER CHURN PREDICTION FOR SYRIATEL: LEVERAGING PREDICTIVE ANALYTICS TO RETAIN CUSTOMERS



PHASE 3 PROJECT: ANGUISTA KUPEKA.

OVERVIEW.

- Customer churn is a significant challenge in the telecommunications industry, where companies like SyriaTel face financial losses due to customers discontinuing their services.
- This project focuses on developing a binary classification model to predict whether a customer is likely to churn.
- The insights derived from this model will empower SyriaTel to proactively identify at-risk customers, implement targeted retention strategies, reduce churn rates, and improve long-term revenue.

BUSINESS UNDERSTANDING.

Objective.

- SyriaTel is experiencing revenue losses due to customer attrition.
- Retaining existing customers is more cost-effective than acquiring new ones.
- The primary objective of this project is to develop a binary classification model to predict customer churn.

Objective cont..

- The model will allow SyriaTel to:

1. Identify customers at risk of churning early.
2. Design and implement targeted retention strategies, such as personalized offers or loyalty programs.
3. Enhance customer satisfaction and loyalty.
4. Minimize revenue losses caused by churn.

Key Business Questions.

1. What are the main factors influencing customer churn?
2. How can these factors be leveraged to design effective retention strategies?
3. How accurately can the model predict customer churn, enabling timely interventions?

DATA UNDERSTANDING.

Dataset Overview.

- The dataset, sourced from Kaggle, contains historical information on SyriaTel customers, covering demographics, usage patterns, account details, support interactions and churn status.

- The dataset consists of 20 columns and 3333 rows.
- Key components include:

Dataset Overview cont..

1. **Target Variable**:
    1. **Churn**: A binary variable indicating whether a customer has churned (1) or not (0).
2. **Predictor Variables**:

- **customer_demographics** = ['account length'] service_usage_metrics = ['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']
- **account_information** = ['area code', 'phone number', 'international plan', 'voice mail plan', 'number vmail messages']
- **customer_support_interactions** = ['customer service calls']

Initial Observations.

- **Class imbalance**: Fewer churn cases compared to non-churn cases, requiring techniques like class weighting or resampling during modeling.
- **Mixed data types**: Numerical and categorical features require preprocessing for compatibility with machine learning models.
- **No missing values detected**, but potential outliers exist in numerical features.

Key Questions for Analysis.

- Are specific usage patterns (e.g., high international call charges) strongly correlated with churn?
- How do categorical variables like the presence of an international plan or a voicemail plan influence churn?
- What role do customer service interactions play in churn prediction?

DATA PREPARATION.

Objective.

- To ensure the dataset is clean, transformed, and ready for modeling by addressing quality issues and engineering relevant features to ensure compatibility with the chosen models and improve prediction accuracy.

Data Cleaning:

1. Handle missing numerical values with imputation (none detected in this dataset).
2. Fill missing categorical values (if present) with 'Unknown'.
3. Remove duplicate records to avoid bias.

Outlier Treatment:

- Detect outliers using the Interquartile Range (IQR) and box plots.
- Apply capping or winsorization to handle extreme values.

Feature Transformation:

1. Standardize numerical features using a StandardScaler for models sensitive to scale.
2. Encode categorical variables using binary encoding.

Feature Engineering:

- Create new features like:

- **Total call duration**: Sum of day, evening, and night call durations.
- **Average call duration per day**: Total call duration divided by the number of days in the account length.
- **Customer service call frequency**: Number of calls normalized by account length.

Class Imbalance Handling:

- Adjusting class weights or winsorization.

Data Splitting.

- Split the dataset into training (75%) and test (25%) subsets for evaluation.

MODELING.

Objective.

- Develop and compare machine learning models to predict customer churn, starting with a base model and progressing to more advanced techniques.

Model Selection:

- **Baseline Model**: Logistic Regression for simplicity and interpretability.
- **Advanced Models**: Decision Tree and Random Forest to capture non-linear patterns and improve accuracy.

Model Training:

- Train each model on the training dataset.
- Apply cross-validation to tune hyperparameters and avoid overfitting.

Evaluation Metrics:

- **Primary Metric**: Accuracy to assess overall performance.
- **Complementary Metrics**: Precision, recall, F1-score, and ROC-AUC to provide a comprehensive evaluation, particularly for handling class imbalance.

EVALUATION.

Objective.

- To assess the model’s performance and ensure it meets business requirements.

Evaluate Test Set Performance:

- Measure accuracy, precision, recall, F1-score, and ROC-AUC.
- Compare test performance with cross-validation results to check for overfitting or underfitting.

Feature Importance Analysis:

- For Logistic Regression, analyze feature coefficients to interpret their impact.
- For Decision Tree and Random Forest, extract feature importance scores to identify key predictors.

Model Comparison:

- Compare the performance of all models to identify the best-performing one.
- Random Forest is expected to outperform in terms of accuracy and AUC.

Business Impact Analysis:

- Use model insights to identify at-risk customers for retention campaigns.
- Quantify potential cost savings from reduced churn rates.

Final Model Deployment:

- Deploy the best-performing model (Random Forest) in a production environment.
- Monitor model performance over time and retrain as needed.

Summary and justification of the Machine Learning Models used: Logistic Regression, Decision Trees and Random Forest.

- After evaluating the performance of the models, the **Random Forest** clearly stands out as the best choice due to its superior overall performance.
- With the highest accuracy (94.12%) and AUC-ROC (0.93), it demonstrates an exceptional ability to differentiate between churn and non-churn customers.
- Notably, the model achieves a precision of 0.97 for predicting churn (Class 1), ensuring that customers identified as likely to churn are highly accurate.
- Although its recall for churn predictions (0.62) is moderate, its robustness and high precision make it particularly suitable for scenarios where minimizing false positives is less critical than accurately identifying true churners.

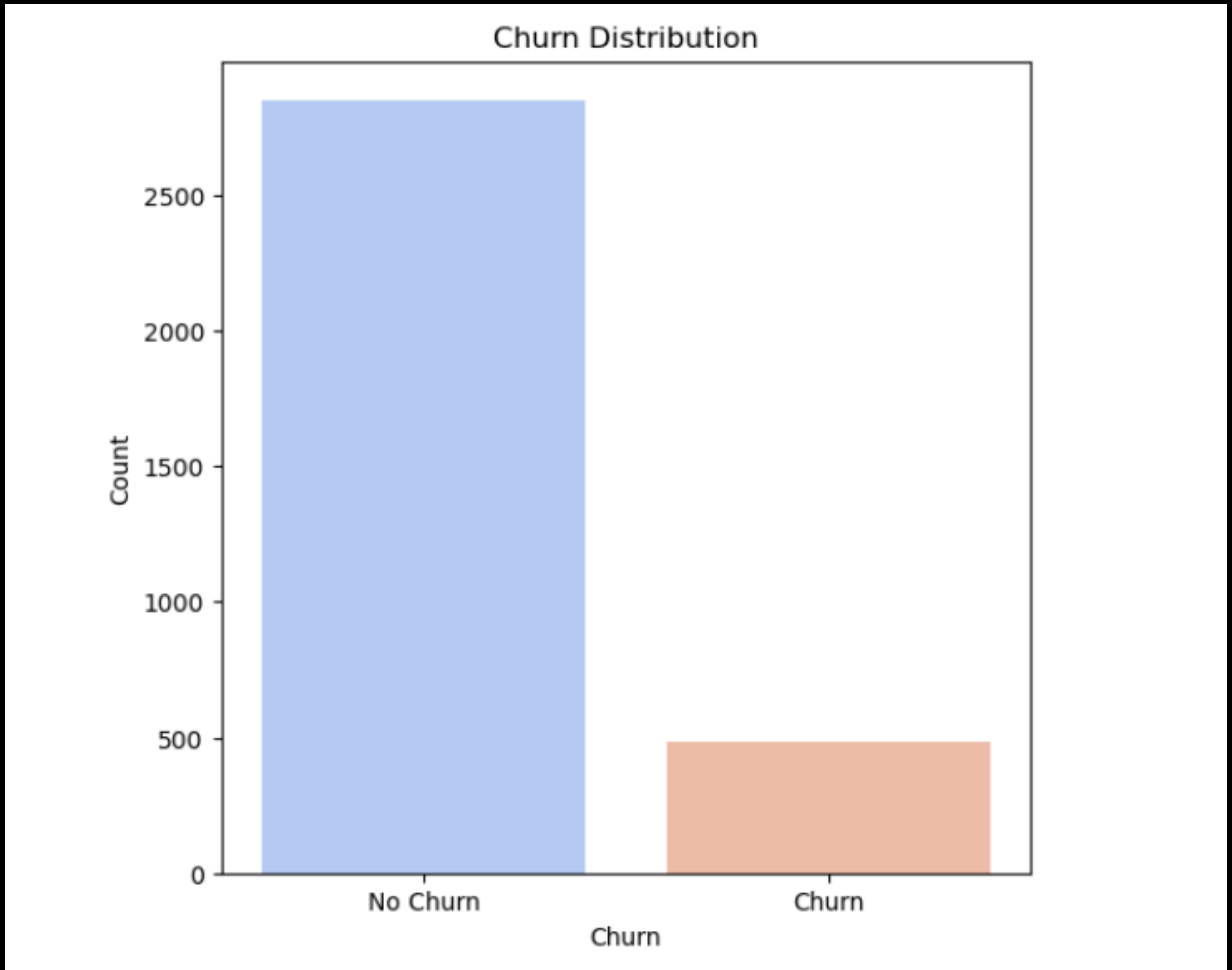
Summary and justification of the Machine Learning Models used: Logistic Regression, Decision Trees and Random Forest cont..

- In comparison, **Logistic Regression** provides simplicity and interpretability but underperforms with a lower accuracy of 80.58% and precision of 0.42 for churn prediction, leading to a higher rate of false positives.
- Similarly, the **Decision Tree** performs better than Logistic Regression in balancing precision and recall; however, its accuracy (90.53%) and AUC-ROC (0.81) remain lower than those of Random Forest.
- Additionally, Decision Trees are susceptible to overfitting, which could impact their reliability without extensive tuning.

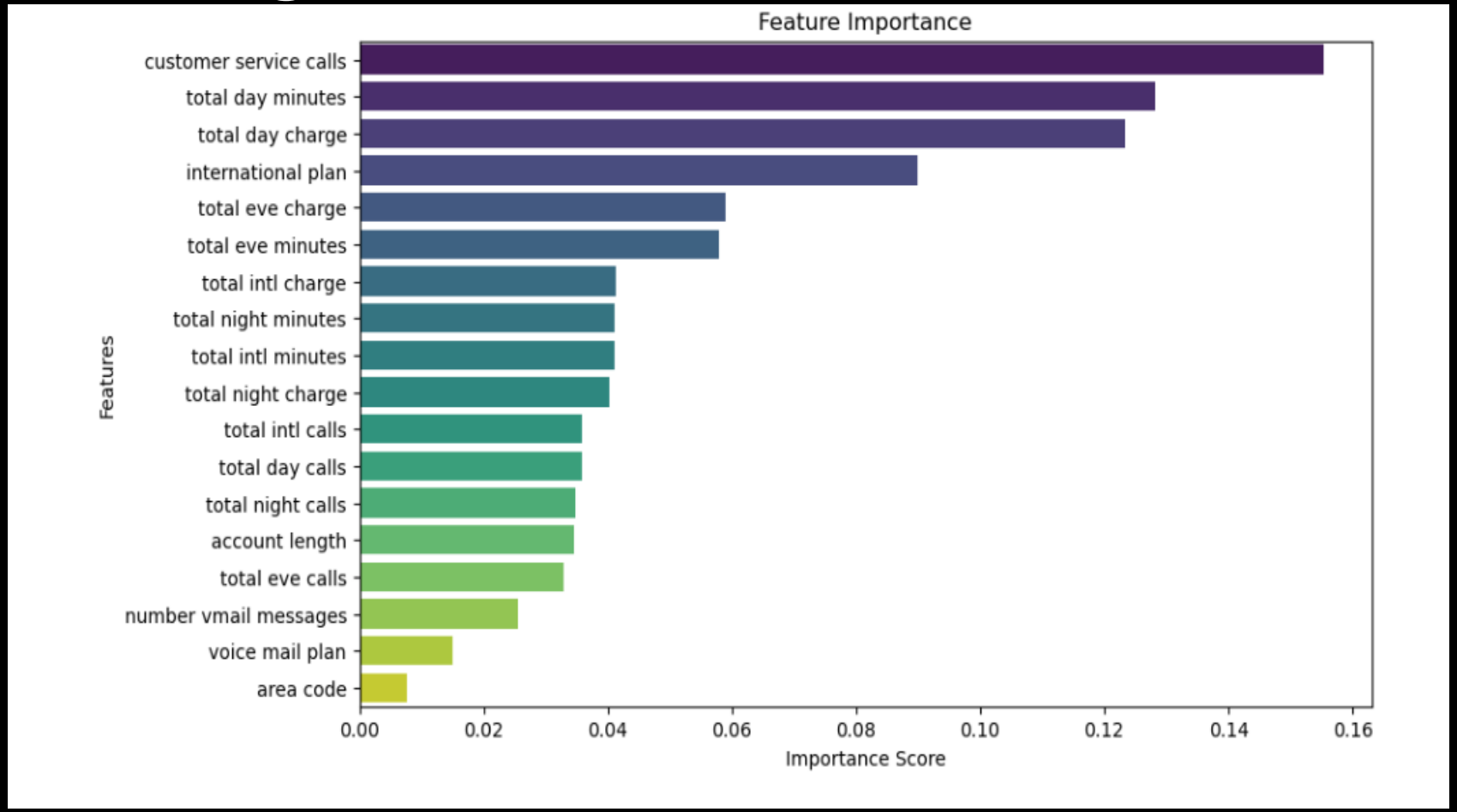
Summary and justification of the Machine Learning Models used: Logistic Regression, Decision Trees and Random Forest cont..

- Considering the business goal of reducing customer churn, the **Random Forest** model is the most reliable and effective option.
- Its ability to maintain a strong balance between precision, recall, and accuracy makes it an ideal candidate for deployment, ensuring that actionable insights can be derived with confidence while minimizing the risk of misclassification.

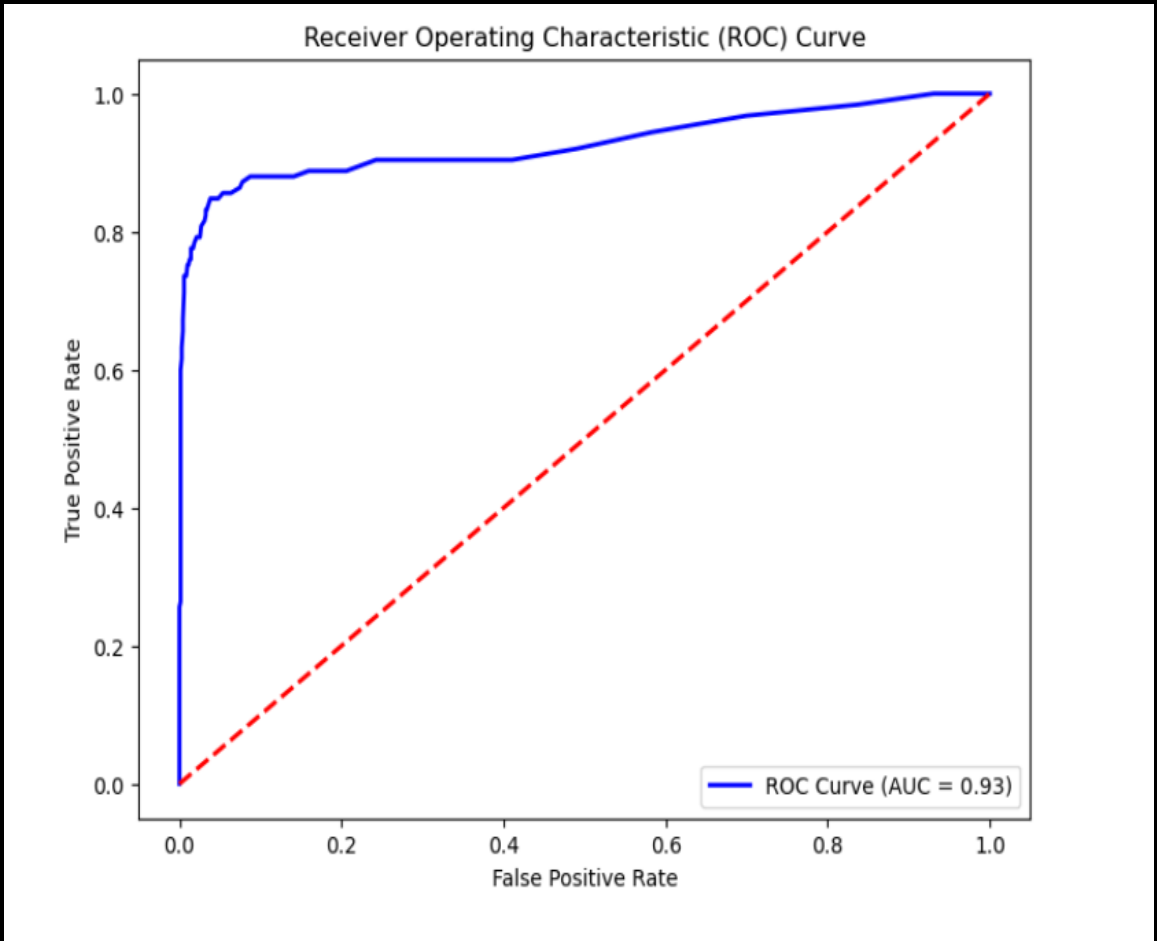
. **Churn Distribution:** Overview of churned vs. non-churned customers.



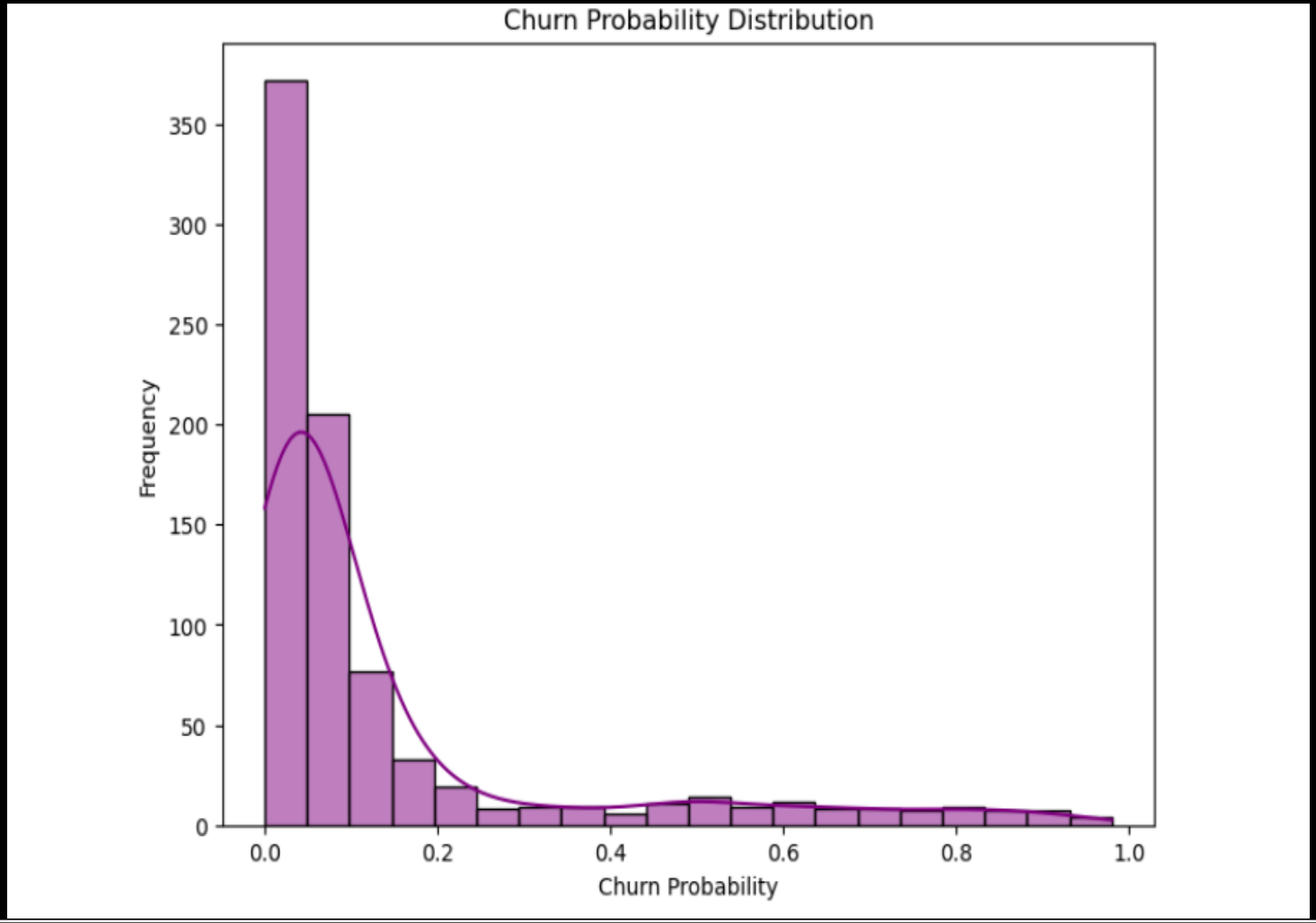
Feature Importance: Highlights key factors contributing to churn.



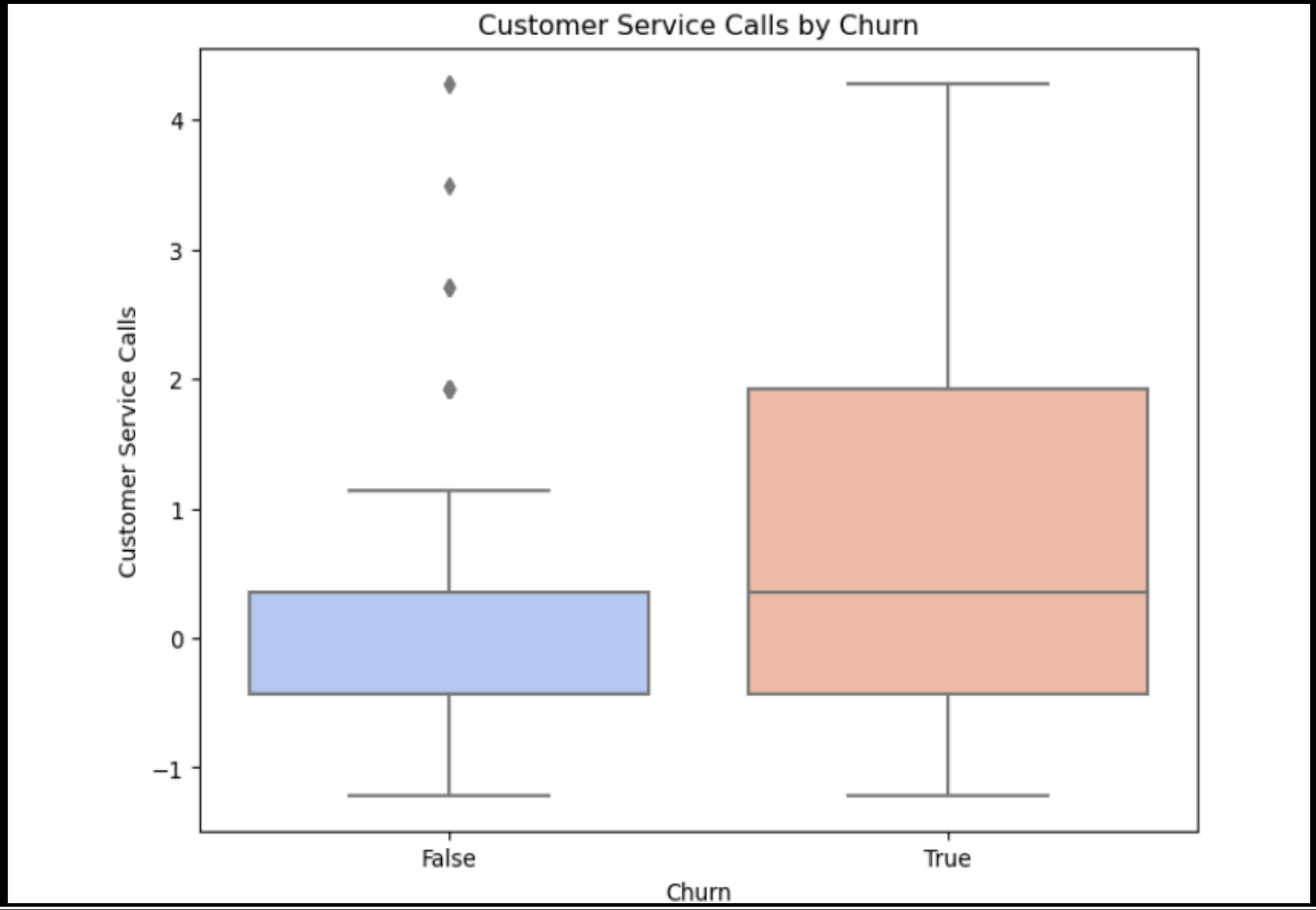
ROC Curve: Demonstrates model performance in distinguishing churners.



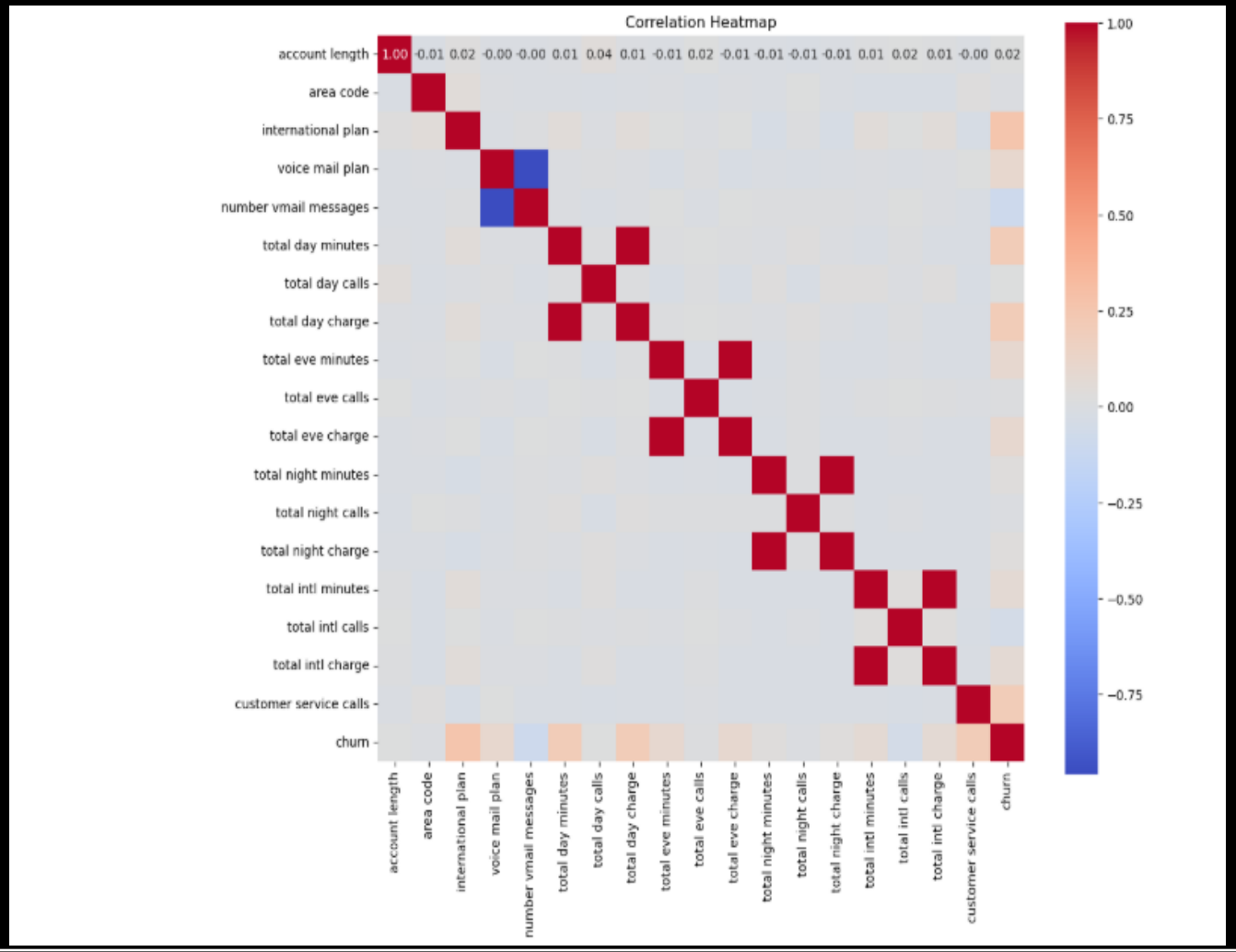
Churn Probability Distribution: Visualizes predictions for all customers.



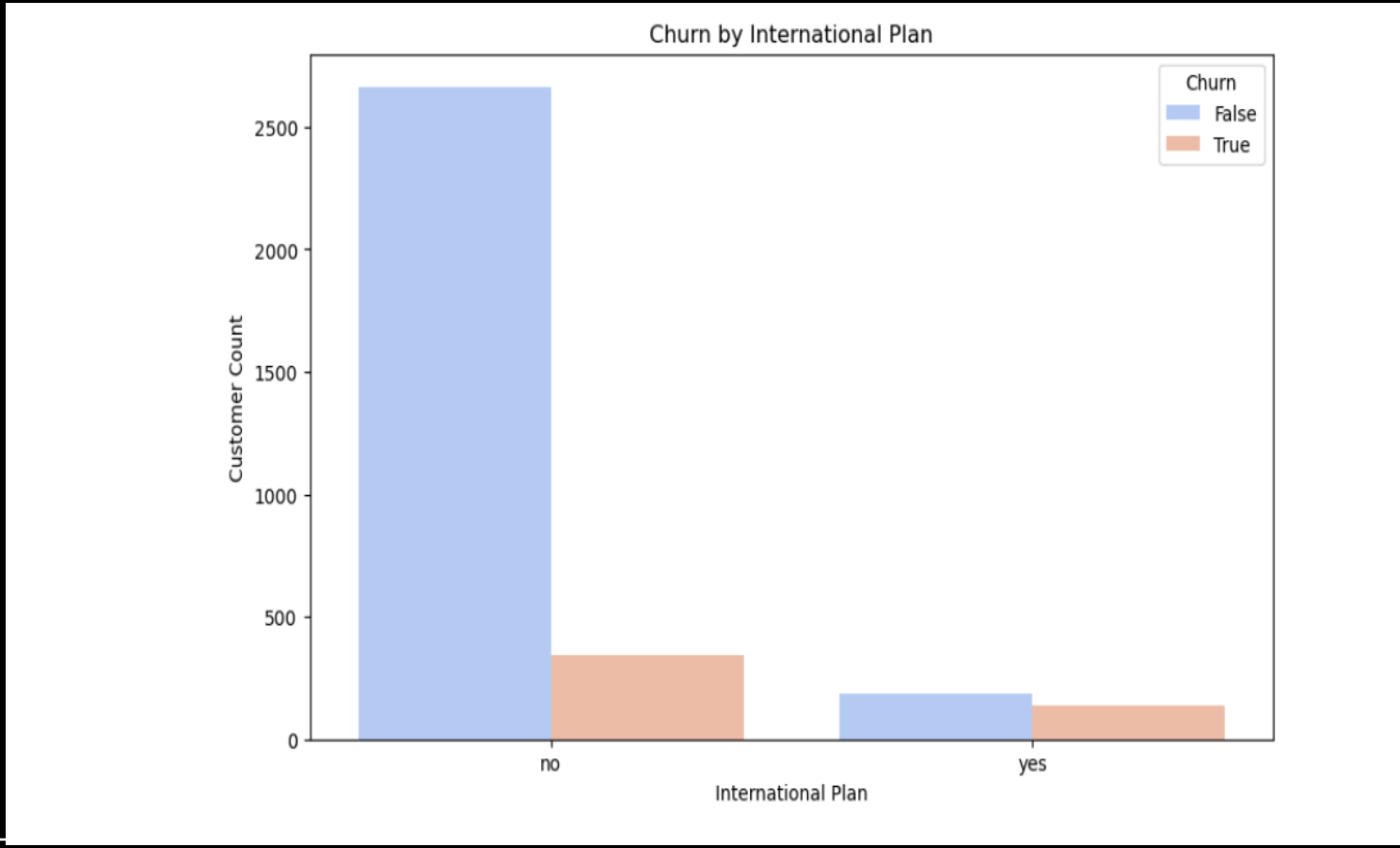
High-Risk Customer Segmentation: Highlights high-risk customers based on key features.



Correlation Heatmap: Shows relationships between features and churn.



Churn by International Plan and Voice Mail Plan: Focuses on churn patterns in plans like the International Plan and Voice Mail Plan



CONCLUSION.

- The deployed Random Forest model demonstrates strong performance in predicting customer churn, as indicated by the evaluation metrics.

- The model achieved an **accuracy of 94%**, reflecting its ability to correctly classify both churn and non-churn customers in the majority of cases. A **precision score of 97%** shows that when the model predicts a customer will churn, it is highly likely to be correct. However, the **recall score of 62%** indicates that the model identified 62% of actual churn cases, leaving room for improvement in capturing more customers at risk of leaving. The **AUC-ROC score of 0.93** highlights the model's excellent ability to distinguish between churners and non-churners.

- The analysis identified **38 high-risk customers** with a churn probability greater than 70%, providing actionable insights for the business. These high-risk customers have been saved to a file (high_risk_customers.csv) for further use by the retention team.

RECOMMENDATIONS.

Retention Strategies:

- Focus on the identified high-risk customers with personalized retention strategies. For instance, targeted campaigns, loyalty rewards, or discounts can help retain these customers.
- Conduct surveys or feedback sessions with high-risk customers to understand their pain points and address their concerns.

Improve Recall:

- Although the model is highly precise, improving recall can help capture a greater proportion of actual churners. Consider experimenting with:
- Adjusting the decision threshold to balance precision and recall.
- Incorporating additional features that may contribute to predicting churn.
- Using ensemble techniques or hyperparameter tuning to enhance the model's performance.

Monitor Model Performance:

Regularly evaluate the model's performance with updated data to ensure it remains effective over time.

Use feedback from retention efforts to refine the model and incorporate new insights into future training.

Expand Efforts Beyond High-Risk Customers:

Develop proactive retention campaigns for medium-risk customers, as addressing potential churn before it escalates can also improve customer retention rates.

Business Impact:

- Leverage the list of high-risk customers to estimate the potential revenue loss and savings from retention campaigns. Use this to justify investment in retention efforts and continuous model improvement.

- NB: By implementing these recommendations, the business can effectively reduce churn rates and improve customer satisfaction, ultimately enhancing overall revenue and brand loyalty.

THANK YOU!

Contact information: anguista.Kupeka@gmail.com