Contoso

SYRIATEL CUSTOMER CHURN PREDICTION

MORINGA SCHOOL

INTRODUCTION

PROJECT OVERVIEW

The SyriaTel Customer Churn project aims to address the high rate of customer attrition within the telecommunications sector. Leveraging customer data such as service plans, call activity, and account details, we will build robust classification models to predict whether a customer is likely to churn. By understanding these patterns, SyriaTel can proactively enhance customer retention strategies, improve customer satisfaction, and reduce financial losses.

Contoso

STAKEHOLDERS

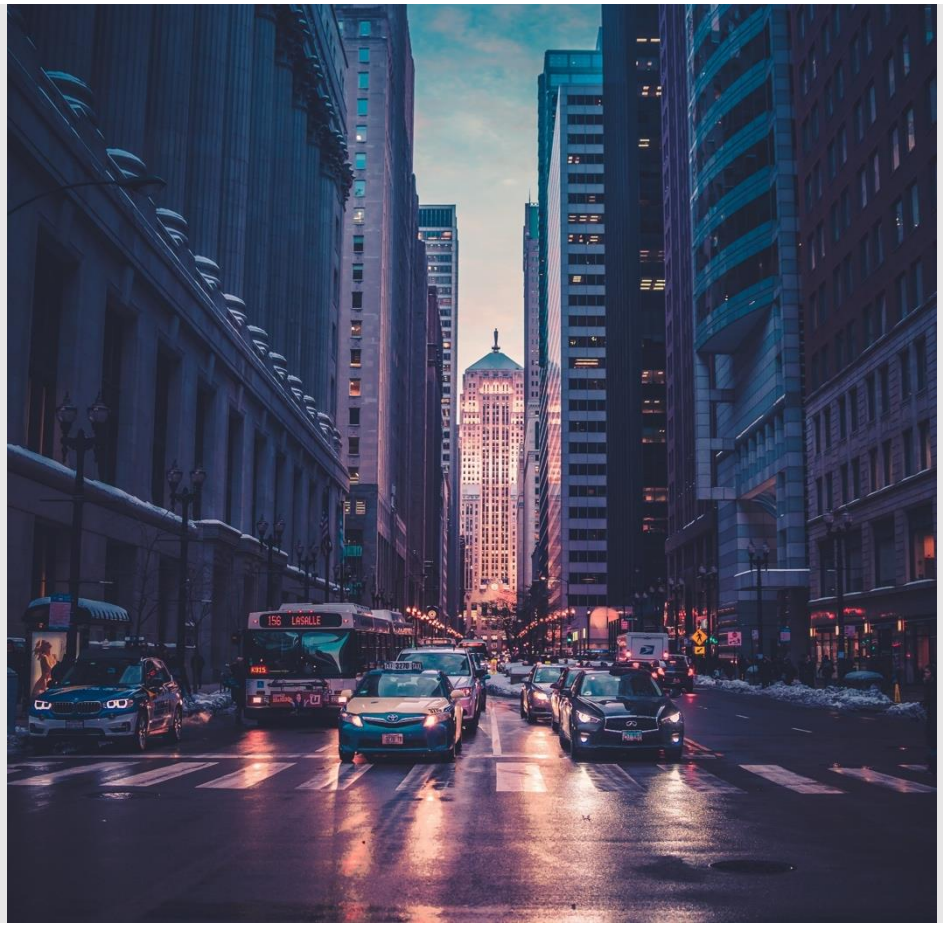
- primary stakeholders for this project include:
- 1. Customer Retention Team: Interested in understanding the factors leading to customer churn in order to develop effective strategies to retain customers.
- 2. Upper Management: Focused on reducing churn rates to improve profitability and customer satisfaction, which are critical to the long-term success of the company.

Contoso

BUSINESS UNDERSTANDING

BUSINESS UNDERSTANDING

SyriaTel is experiencing high customer churn, negatively affecting revenue and growth. The company seeks to understand the factors contributing to churn and develop a predictive model to identify at-risk customers. By predicting churn, SyriaTel can intervene with targeted retention strategies, ensuring a more stable customer base.



PROBLEM STATEMENT

SyriaTel is losing customers at an alarming rate, impacting its revenue and market position. The company needs a data-driven approach to understand churn dynamics, enabling proactive measures to retain customers. The primary goal is to build a model that can accurately predict customer churn and provide actionable insights to improve retention strategies.

Contoso

DATA DESCRIPTION

COLUMN DESCRIPTIONS

State: The customer's location, represented as a categorical variable.

Account Length: Duration of the customer's account in days.

Area Code: Numeric representation of the customer's area code.

International Plan: Whether the customer has subscribed to an international calling plan.

).

Voice Mail Plan: Subscription status to a voicemail plan.

Total Day/Eve/Night/Intl Minutes: Usage minutes during different time segments.

Total Day/Eve/Night/Intl Calls: Call counts across different time segments.

Total Day/Eve/Night/Intl Charge: Charges accrued in different time segments.

Customer Service Calls: Number of calls made to customer service.

Churn: The target variable indicating whether the customer has churned (True/False

SPECIFIC OBJECTIVES

OBJECTIVES

1. *Identify Key Determinants of Customer Churn*
2. *Model Selection and Performance Evaluation:*
3. *Provide Recommendations for Retention Strategies:*

Contoso

METHODOLOGY

DATA UNDERSTANDING AND PREPARATION

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	tota nigh minute
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	...	99	16.78	244.
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	...	103	16.62	254.
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	...	110	10.30	162.
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	...	88	5.26	196.
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	...	122	12.61	186.

5 rows × 21 columns

- Load the dataset and understand the Dataset Features.
- The Dataset has 3333 Rows and 21 columns.
- It is made of 1 bool type, 8 integers, 8 float type, and 4 object type

Contoso

DATA UNDERSTANDING AND PREPARATION

- Our target variable is a bool type
- The categorical variables are *state, phone number, international plan, voice mail plan.*
- There is 17 numericals variables which are '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'

Contoso

EXPLORATORY DATA ANALYSIS (EDA)

- Perform EDA to understand data distribution and correlations.
    - *Exploratory Data Analysis (EDA) is the process of examining a dataset's main features using statistical tools and visualizations. It helps identify patterns, spot anomalies, and understand relationships between variables. EDA provides a clear overview of the data, guiding further analysis and model selection.*

Contoso

SPLITTING THE DATASET INTO CATEGORICAL AND NUMERICAL VARIABLES.

- Will start with the Categorical Variables of our dataset, and check the features in our categorical dataset.

Categorical Variables

	state	phone number	international plan	voice mail plan
0	KS	382-4657	no	yes
1	OH	371-7191	no	yes
2	NJ	358-1921	no	no
3	OH	375-9999	yes	no
4	OK	330-6626	yes	no

Contoso

SPLITTING THE DATASET INTO CATEGORICAL AND NUMERICAL VARIABLES.

- Then to the numerical variables

Numerical Variables

	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
0	128	415	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1
1	107	415	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1
2	137	415	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0
3	84	408	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2
4	75	415	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3

Contoso

EXPLORING PROBLEMS WITHIN OUR VARIABLES

CATEGORICAL VARIABLES

a) Checking for Null values

- Our Categorical Variables had no Missing Values.

b) Checked for Cardinality

- Cardinality refers to the number of unique values or labels in a categorical variable. It helps to understand the diversity or variety within that variable. High cardinality means the variable has many unique values, while low cardinality means it has fewer unique values.
- I found the results as follows:

➢ state contains 51 labels
➢ phone number contains 3333 labels
➢ international plan contains 2 labels
➢ voice mail plan contains 2 labels

Contoso

EXPLORING PROBLEMS WITHIN OUR VARIABLES

CATEGORICAL VARIABLES

- Phone number has very high cardinality which might not be useful for the model. Phone numbers are typically unique identifiers and don't provide meaningful information for prediction. I might consider dropping it. So unfortunately we had to drop the phone number column here simply because it had high cardinality and it would have affected our model.

Contoso

EXPLORING PROBLEMS WITHIN OUR VARIABLES

CATEGORICAL VARIABLES

- ***Exploring the State Variable***
  - State contains 51 labels
  - Checking Frequency distribution of our Values:
    - `state`
    - `WV =106,MN=84,NY =83,AL=80,WI=78, OH=78,OR=78,WY=77,VA=77,CT=74,MI=73,ID =73,VT=73,TX=72,UT=72,IN=71,MD=70,KS=7 0,NC=68,NJ=68,MT=68,CO=66,NV=66,WA=66, LA=51,PA=45,IA=44,CA=34`
- ***Exploring International Plan Variable***
  - International plan contains 2 labels
  - Checking the labels :
    - `international plan`
    - `no 3010`
    - `yes 323`

Contoso

EXPLORING PROBLEMS WITHIN OUR VARIABLES

CATEGORICAL VARIABLES

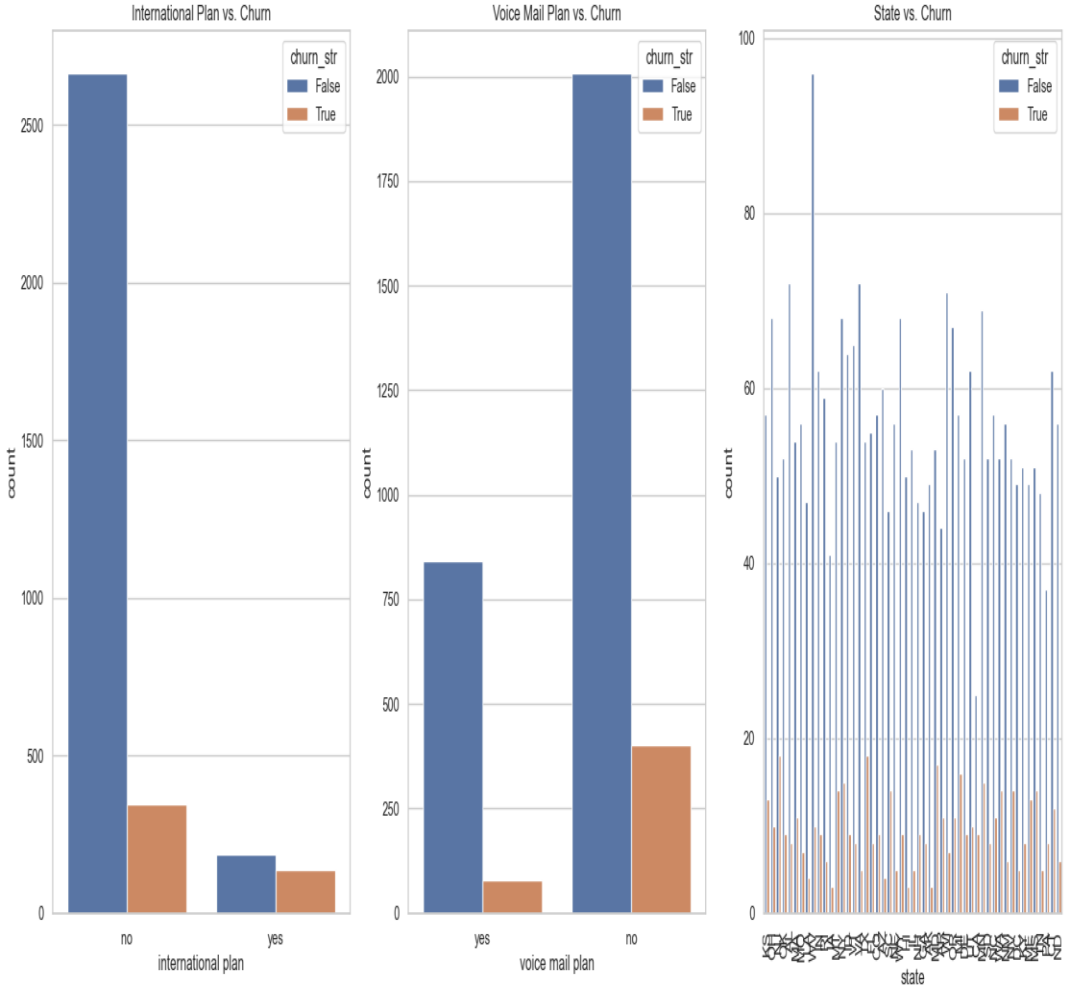
- ***Exploring Voice Mail Plan Variable***

➢ voice mail plan contains 2 labels

➢ Checking the labels :

```
voice mail plan
no      2411
yes      922
```

COUNTPLOTS FOR CATEGORICAL VS CHURN



INTERPRETATION OF THE PLOTS

- **Interpretation of the Plots**
- ***International Plan vs. Churn:***
- Customers without the international plan are significantly less likely to churn compared to those with the plan.
- A higher proportion of customers with the international plan churn compared to those without it, indicating that having an international plan is associated with higher churn rates.
- ***Voice Mail Plan vs. Churn:***
- Customers with the voice mail plan are less likely to churn compared to those without it.
- The churn rate is higher among customers who do not have the voice mail plan, suggesting that not having this plan might be associated with increased churn.

Contoso

INTERPRETATION OF THE PLOTS

- ***State vs. Churn:***
- The churn rates are fairly consistent across different states, with no particular state showing an extremely high or low churn rate compared to others.
- The variation in churn rates across states does not seem substantial, suggesting that the state variable might not be a strong predictor of churn.

- **Conclusion**
- State Variable: Given the visual analysis, the state does not appear to have a significant impact on chur as there is no clear pattern or significant differences i churn rates across states. Dropping this variable might be appropriate inorder to simplify the model.
- International and Voice Mail Plans: These variables show clear differences in churn rates, making them important features to include in your analysis or predictive model.
- .

Contoso

EXPLORING NUMERICAL VARIABLES

- There are 16 numerical variables
- The numerical variables are : ['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', 'tot intl minutes', 'total intl calls', 'tota intl charge', 'customer service calls']
-
- There are no Missing Values in our Numerical data.
- Checking The Descriptive statics of our Numerical data
- .

EXPLORING NUMERICAL VARIABLES

- .

	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 nig chan
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.0000
mean	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.0393
std	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.2758
min	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.0400
25%	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.5200
50%	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.0500
75%	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.5900
max	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.7700

Contoso

CHECKING FOR OUTLIERS

THERE ARE AS FOLLOWS :

- ACCOUNT LENGTH HAS 7 OUTLIERS.
- TOTAL DAY MINUTES HAS 9 OUTLIERS.
- TOTAL DAY CALLS HAS 9 OUTLIERS.
- TOTAL DAY CHARGE HAS 9 OUTLIERS.
- TOTAL EVE MINUTES HAS 9 OUTLIERS.
- TOTAL EVE CALLS HAS 7 OUTLIERS.
- TOTAL EVE CHARGE HAS 9 OUTLIERS.
- TOTAL NIGHT MINUTES HAS 11 OUTLIERS.
- TOTAL NIGHT CALLS HAS 6 OUTLIERS.
- TOTAL NIGHT CHARGE HAS 11 OUTLIERS.
- TOTAL INTL MINUTES HAS 22 OUTLIERS.
- TOTAL INTL CHARGE HAS 22 OUTLIERS

Contoso

- .

HANDLING OUTLIERS

- There's some outliers in our variables which can affect our model we will find techniques of tackling them.
- But before tackling our problems we have to check distribution of our data if they are skewed or normalized so if they normal I do extreme Value analysis, and if skewed I find the IQR.
- I created a function which showed if it was normally or skewed

Contoso

Declare Feature Vector and Target Variable

```
X = df.drop('churn', axis=1)
y = df['churn']
```

✓ 0.0s Python

Then we split the data into separate training and test set

```
# split X and y into training and testing sets

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

✓ 1.3s Python

Contoso

BASELINE MODEL: LOGISTIC REGRESSION:

- Modelling of Logistic regression model and predicting the results, the baseline model had an accuracy score of 87.26%, indicating the model correctly classified about 87% of test data.
- We had also to check for overfitting and underfitting:
- Training set score: 0.8620
- Test set score: 0.8786
- The training set accuracy is 0.8597 while the test s accuracy is 0.8726. we have used the default value C= which provides an accuracy of 87 on the test data and on the training set which is coparable, I will increase C and fit a more flexible model and check the score
- I tried by changing the C to 100 and 0.01 and got the best model after hyperparameter tuning as Best parameters for Logistic Regression: {'C': 10, 'solver': 'liblinear'}

MORE COMPLEX MODELS

- We also did decision tree and random forest models

Which we got more better and performing model than our logistic regression and did a stacking ensembled model consisting the best of the three logistic, decision and random forest models

Contoso

STACKING MODEL CLASSIFICATION REPORT:

- *precision recall f1-score support*
-
- *0 0.98 0.97 0.97 579*
- *1 0.83 0.84 0.84 88*
-
- *accuracy 0.96 667*
- *macro avg 0.90 0.91 0.91 667*
- *weighted avg 0.96 0.96 0.96 667*
-
- *Confusion Matrix:*
- *[[564 15]*
- *[14 74]]*
- *Stacking Model AUC Score: 0.9227311979902654*

Contoso

STACKING MODEL CLASSIFICATION REPORT:

- Stacking Ensemble significantly improves the performance of the Random Forest model, making it the most suitable choice for the business objective of reducing customer loss. The stacking model outperforms the individual models and achieves a high AUC score, indicating better discriminative power. Further improvements can be explored through advanced ensemble techniques and fine-tuning.

- Performance: The stacking model shows excellent performance across all metrics. It achieves a high accuracy (96%) and a very high AUC score (0.9269), indicating strong discriminative power.

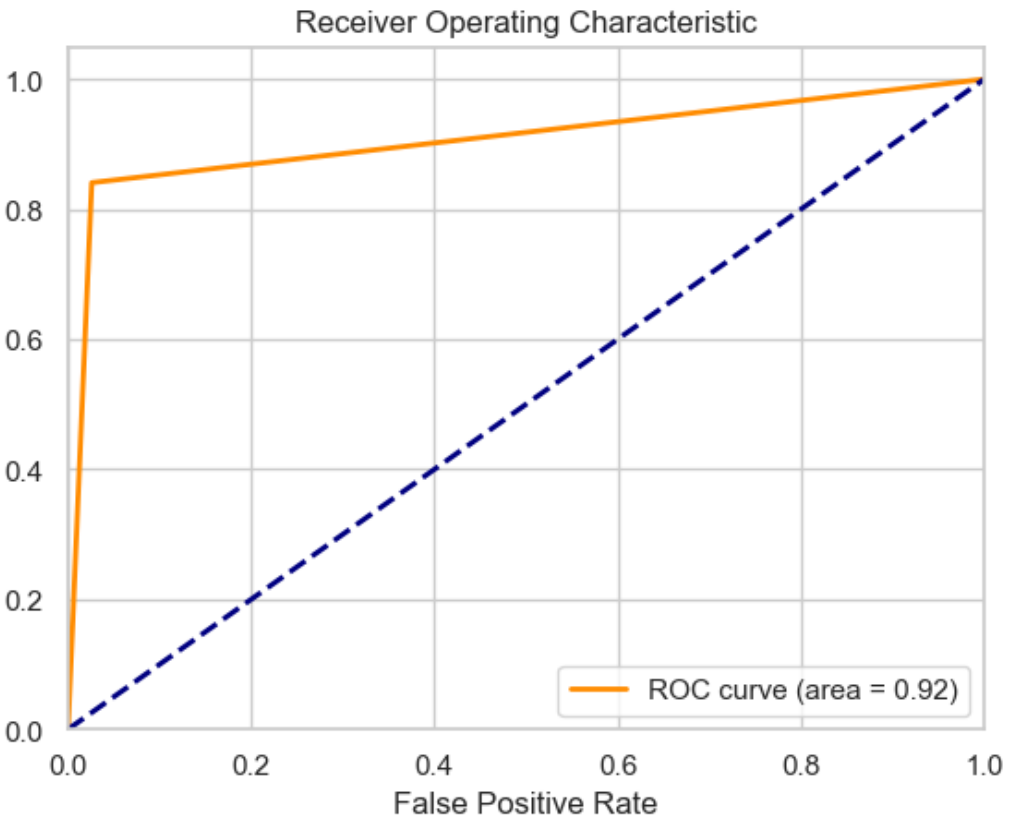
- Class 1 (Churn) Performance: Precision and recall for churn (Class 1) are both improved compared to the individual models, suggesting that stacking effectively combines their strengths.

- Confusion Matrix: The stacking model reduces false positives and false negatives, with better balance in predicting both classes compared to individual models

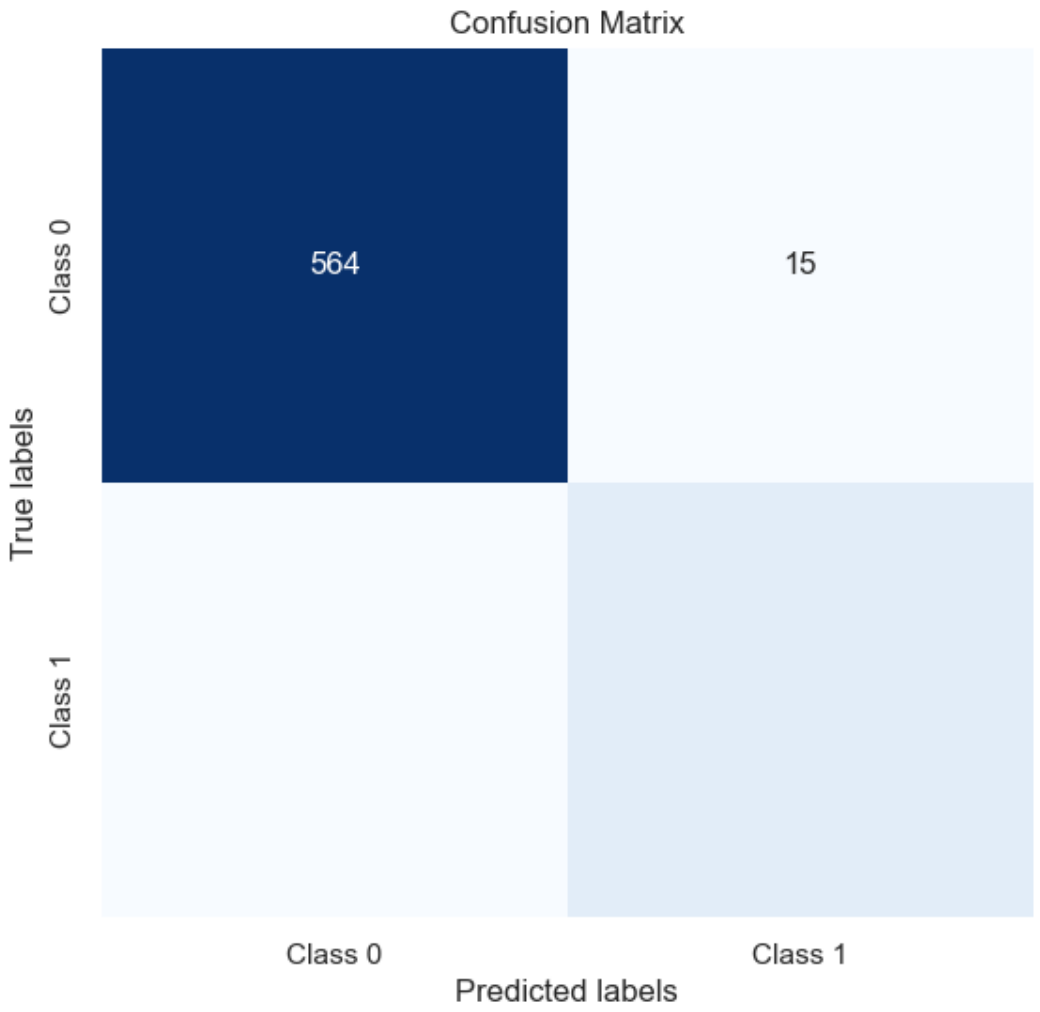
Contoso

ROC CURVE OF THE STACKING ENSEMBLING MODEL

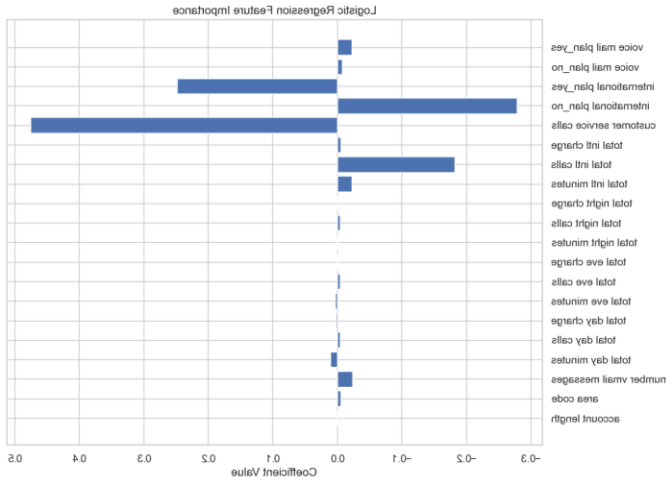
:



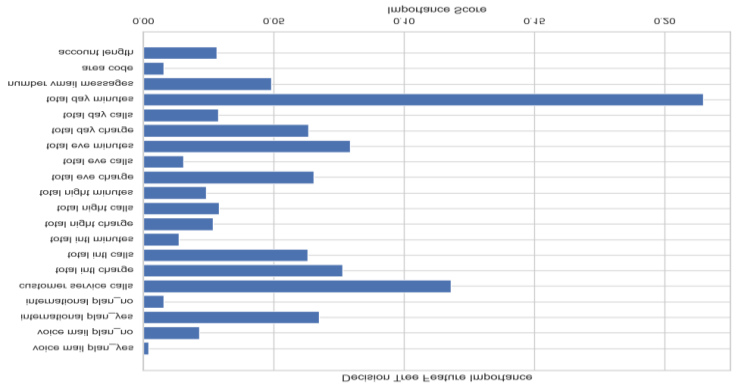
CONFUSION MATRIX FOR STACKING ENSEMBLING MODEL



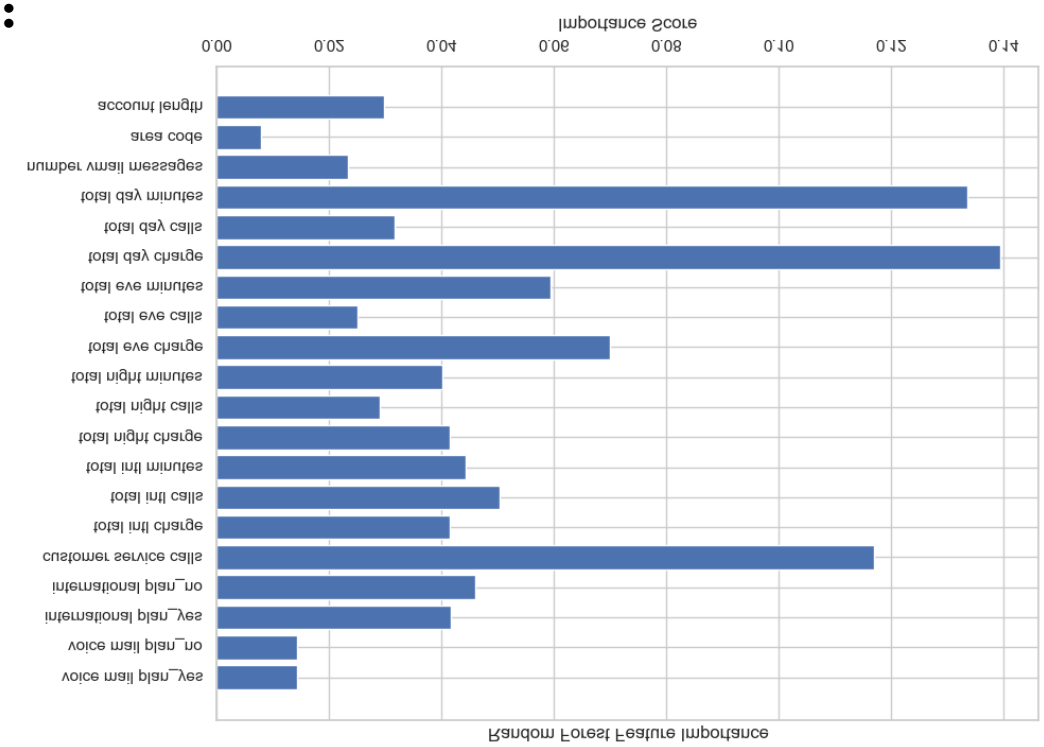
FEATURE IMPORTANCE
FOR LOGISTIC REGRESSION



FOR DECISION TREE



RANDOM FOREST



FEATURE IMPORTANCE:

- The feature importance scores provided indicate the significance of each feature in predicting the target variable in the random forest classifier. Here's an interpretation of the key points:

- `Total Day Minutes (0.15072)` and `Total Day Charge (0.13237)` are the most influential features in your model. This suggests that the amount of time and associated charges during the day are strong indicators of whether a customer is likely to churn.

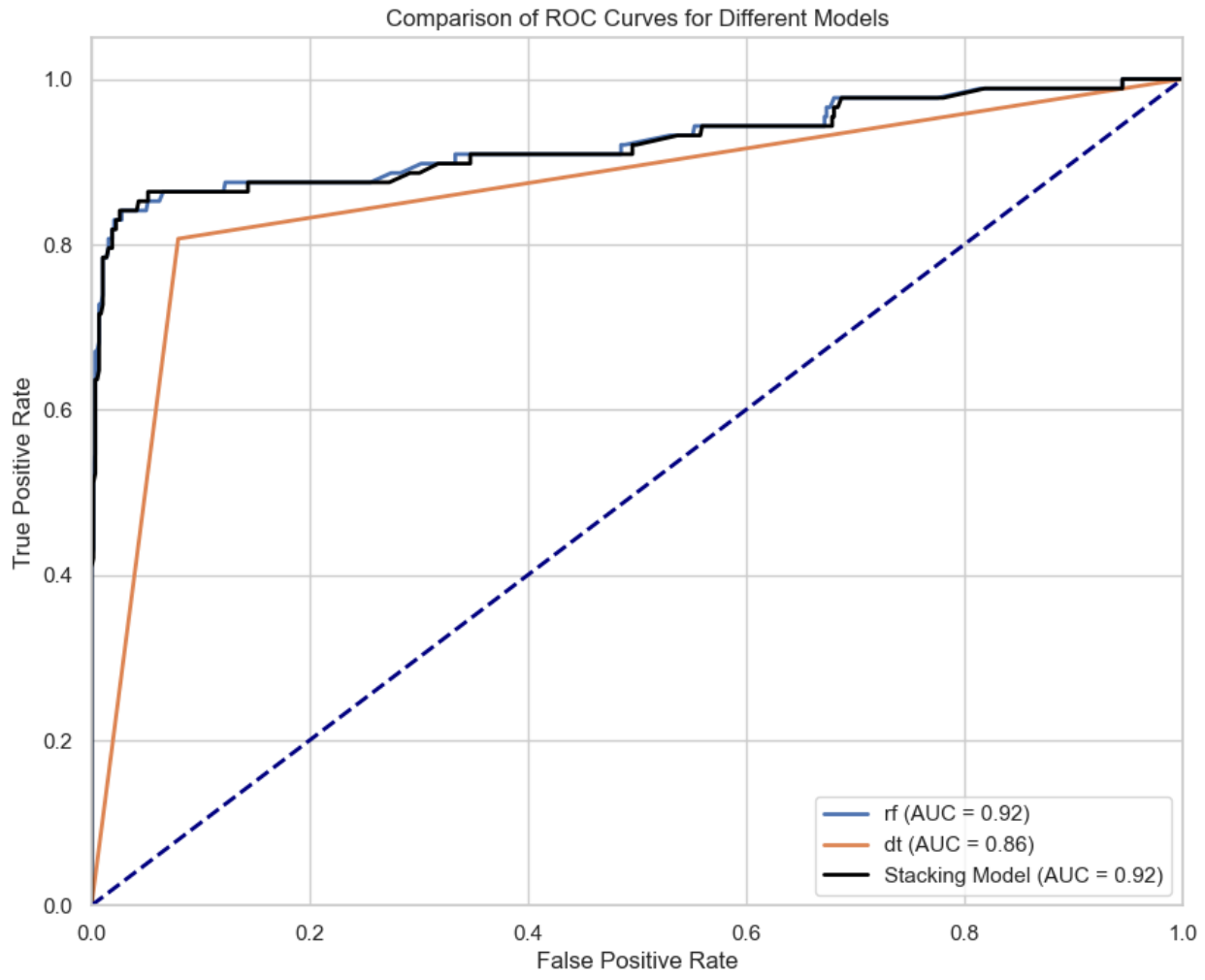
- `Customer Service Calls (0.11478)` also plays a significant role, indicating that customers who contact customer service more frequently may be more likely to churn.

- `Total Intl Calls (0.05300)` and `Total Intl Minutes (0.04715)` suggest that international calling behavior is also a notable factor in predicting churn.

- `Total Eve Minutes (0.06316)` and `Total Eve Charge (0.06375)` are moderately important, indicating that evening usage patterns have some influence on churn.

- Features like `Area Code (0.00673)` and `Voice Mail Plan (0.01342/0.01367)` have lower importance, suggesting they are less influential in predicting customer churn

MODEL COMPARISON OF ROC CURVES



Contoso

CONCLUSION:

- Best Model: The stacking ensemble outperforms both the Decision Tree and Random Forest models in terms of precision, recall, and overall accuracy.

- Recommendation: The stacking model is recommended for deployment due to its superior performance. It effectively leverages the strengths of the base models to improve predictive accuracy for customer churn.

Objective 1: Identify Key Determinants of Customer Churn

- Conclusion:

1. Findings: The analysis identified several key factors influencing customer churn at SyriaTel. Features like Total Day Minutes, Total Day Charge, Customer Service Calls, Total Intl Calls, and the International Plan were among the most significant predictors of churn.

- Specifically, higher total day minutes and charges were strongly associated with an increased likelihood of churn. Additionally, customers with frequent customer service calls or an international plan were more likely to leave, possibly indicating dissatisfaction or unmet needs.

-

- 2. Implication: These findings suggest that customers who heavily use daytime services or interact frequently with customer service are at a higher risk of churn. This could imply dissatisfaction with service quality, pricing, or the perceived value of the services provided. The presence of an international plan as a churn factor indicates that the international calling service might not be meeting customer expectations in terms of cost or quality.

OBJECTIVE 2: MODEL SELECTION AND PERFORMANCE EVALUATION

:

- Conclusion:

- Chosen Model: The stacking ensemble model has been selected as the best-performing model due to its superior predictive accuracy and ability to integrate multiple algorithms to improve overall performance. This model's high ROC-AUC score makes it ideal for deployment in predicting customer churn at SyriaTel.

-

- **Objective 3: Provide Recommendations for Retention Strategies**

- **Conclusion:**

- •Findings: Based on the model's findings, the following strategies are recommended to help SyriaTel retain customers:

- 1.Improve Customer Service Quality: Since high customer service interactions correlate with churn, investing in better training and reducing wait times could enhance customer satisfaction, reducing churn rates.

- 2.Offer Customized Plans: Customers with higher day-time usage might benefit from personalized plans that offer discounts or incentives for heavy usage during peak hours. This could increase loyalty by providing more value to the customers.

Contoso

- 3 .Monitor and Engage High-Risk Customers: Use the model to identify customers at high risk of churn and proactively reach out with retention offers or satisfaction surveys. This can help address issues before customers decide to leave.

- 4.Enhance International Plan Offerings: Since the international plan is a significant churn factor, consider revising the plan's pricing, coverage, or adding new features that make it more attractive to customers.

Recommendations:

- **1.Enhance Customer Support:**
- o Focus: Customers who frequently contact customer service are more likely to churn.
- oAction: Improve the customer service experience by offering more training to support staff, introducing more efficient problem-resolution processes, and potentially using AI-driven customer servi tools to anticipate and address customer issues before they escalate.

- **2. Revise International Plans:**
- o Focus: International plan subscribers show a higher tendency to churn.
- o Action: Reevaluate and enhance international calling plans to offer better value, such as reducing rates or bundling with other services. Consider introducing targeted promotions or discounts for international callers to increase satisfaction and loyalty.
-

- **3. Optimize Daytime Service Plans:**
- o Focus: High daytime usage is strongly correlated with churn.
- o Action: Introduce or promote plans that offer better value for heavy daytime users, such as unlimited or higher minute allowances, to prevent these customers from seeking better deals elsewhere.
- **4. Proactive Engagement with High-Risk Customers:**
- o Focus: Use the model to identify at-risk customers early.
- o Action: Implement a proactive retention strategy by contacting these customers before they consider leaving. Offer personalized discounts, loyalty rewards, or check-ins to ensure they feel valued and understood by SyriaTel.
- **5. Continuous Monitoring and Model Refinement:**
- o Focus: Even the best models can be improved over time.
- o Action: Regularly update the model with new data to maintain its accuracy and effectiveness. Additionally, monitor the impact of implemented retention strategies and adjust them based on ongoing feedback and data analysis.
- By focusing on these key areas, SyriaTel can effectively reduce customer churn, improve service quality, and enhance overall customer satisfaction and loyalty.
-

Contoso

Contoso

THANK YOU

VCTRMUTHOKA@GMAIL.COM

HTTPS://GITHUB.COM/VICTORMUUO07/PHASE-3-FINAL-PROJECT.GIT