KEEPING CUSTOMERS CONNECTED - AND NOT DISCONNECTED!

THE SYRIATEL DATA REPORT

1.BUSINESS UNDERSTANDING

1.1 Business Overview

A telecommunications company is an organization that provides services for long distance communication. They do this by building and maintaining the physical networks, like cell towers, that transmit signals to individuals and businesses. These companies facilitate essential services like accessing the internet, making phone calls and sending messages. They make money through customer subscriptions and usage fees for these services. SyriaTel is a telecom company that provides call, text and data services to customers. One advantage of working with this company is that it is a high-performing sector that contributes to economic growth, potentially increasing returns for investors. Telecommunications is also an essential service with steady demand, making it stable and a valuable industry to be part of. However, the telecom industry is highly competitive and customers can easily switch to other providers if they're dissatisfied. This creates a high risk of customer churn, which can reduce revenue and can discourage investor confidence if not properly managed.

1.2 Problem Statement

SyriaTel might lose customers to competitors; by analyzing customer data, we predicted churn and uncovered the reasons why customers leave, so SyriaTel can take action to reduce churn and improve customer retention.

This is costly because:

Revenue loss: Each customer lost means recurring revenue lost.

High acquisition cost: It is more expensive to acquire a new customer than to retain an existing one.

Competitive pressure: In a competitive market, reducing churn is critical for survival and growth.

We predicted the customers that were likely to leave so that SyriaTel could take action early e.g. giving offers, improving services, or solving problems to make those customers stay.

So, the goal was to reduce churn and keep loyal customers.

1.3 Business Objectives

1.3.1 *Main objective:*

We predicted customer churn and provided insights that would help SyriaTel keep its customers and reduce revenue loss.

1.3.2 *Specific objectives:*

1. We developed a model that predicts whether a customer will churn or stay.
2. We identified the key factors e.g. Total day minutes, customer service calls, total day charge, international plan were among the factors that influence the probability of a customer to churn or not to churn.
3. We provided insights that SyriaTel could use to design strategies for reducing churn and improving customer satisfaction.
4. We determined New Jersey as the state with the highest churning rate.

1.3 Success Criteria

Model performance The churn prediction model achieves a good level of balance between recall and ROC-AUC score in order to correctly identifying customers who churn.

Insights gained The analysis clearly identified the key factors that contribute to churn eg. high call charges and frequent complains. ***Business value*** SyriaTel can use the model's results to take practical actions, such as designing loyalty offers or improving customer service which can help improve customer churn.

2. DATA UNDERSTANDING

2.1 Data Understanding

We worked with the Syria Tel customer churn dataset from Kaggle Link, it had a total of 21 columns and 3,333 rows. Below are the descriptions of the columns:

1. **state** – U.S. state where the customer lives.
2. **account length** – Number of days the customer has had the account.
3. **area code** – Telephone area code.
4. **phone number** – Customer’s phone number.
5. **international plan** – Whether the customer has an international calling plan (yes/no).
6. **voice mail plan** – Whether the customer has a voicemail plan (yes/no).
7. **number vmail messages** – Number of voicemail messages the customer has.
8. **total day minutes** – Total minutes of calls during the day.
9. **total day calls** – Number of calls during the day.
10. **total day charge** – Total charges for day calls.
11. **total eve minutes** – Total minutes of calls during the evening.
12. **total eve calls** – Number of calls during the evening.
13. **total eve charge** – Total charges for evening calls.
14. **total night minutes** – Total minutes of calls during the night.
15. **total night calls** – Number of calls during the night.
16. **total night charge** – Total charges for night calls.
17. **total intl minutes** – Total minutes of international calls.
18. **total intl calls** – Number of international calls.

19. **total intl charge** – Total charges for international calls.

20. **customer service calls** – Number of calls made to customer service.

21. **churn** – Target variable: whether the customer left the company (True = churned, False = stayed).

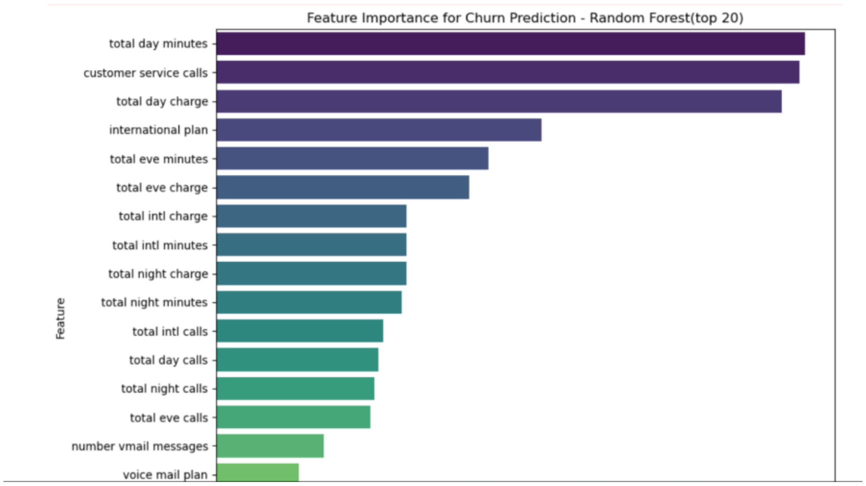
The columns were made up of both categorical (3) and numerical variables (18).

2.2 Data preparation

- The dataset had no missing values.
- The dataset had no duplicates.
- The dataset had no inconsistencies e.g. a mix up of numbers and symbols.
- The dataset had no placeholders.
- The target variable was imbalanced with false being 85.5% and true being 14.49% so we had to correct that.
- The columns that we onehotencoded were; state, churn, voicemail plan and international plan.
- The dataset had some outliers in columns like, total day minutes, total eve minutes, total night minutes and account length. – We decided to drop these columns because they were to affect the model's performance as they also had multicollinearity.

2.3 Data explanation

2.3.1 Feature importance



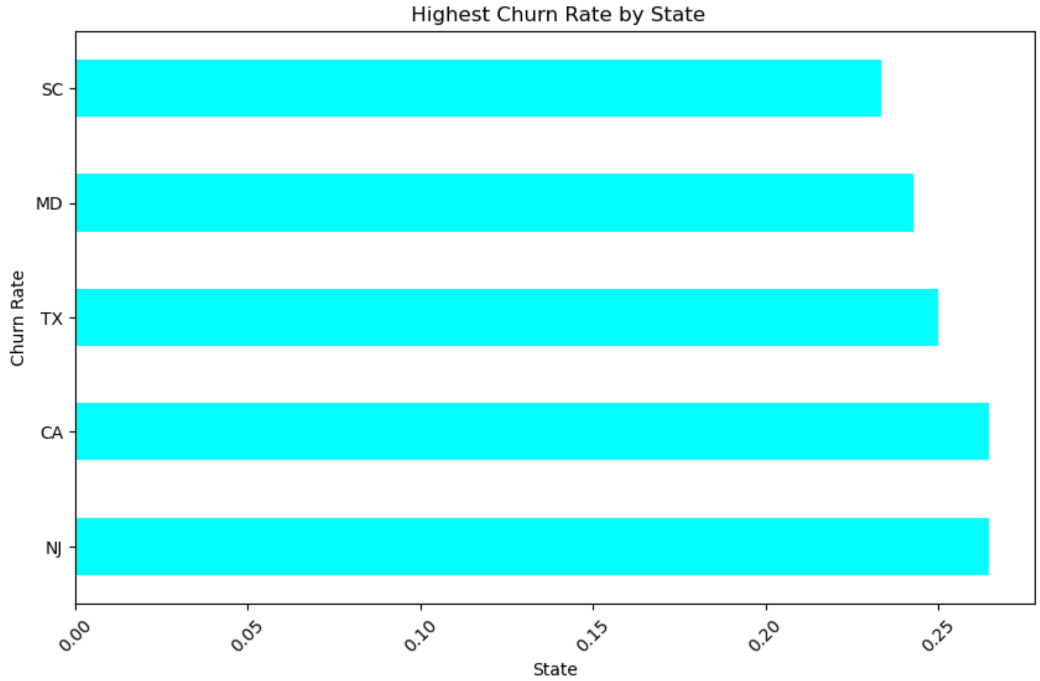
Customer service performance is the strongest churn signal - improving this can directly reduce churn.

High-usage customers (especially daytime callers) are at risk - need targeted retention offers.

International plan users are sensitive - better international packages could keep them loyal.

Voicemail and state-level differences don’t matter much - no need to allocate marketing budget here.

2.3.2 States with the highest churn rates



The bar plot above shows the states with the highest churn rates. The description is as follows:

In **New Jersey (NJ)**, Out of 68 customers, 18 churned giving a 26.5% churn rate.

In **California (CA)**, Out of 34 customers, 9 churned giving a 26.5% churn rate.

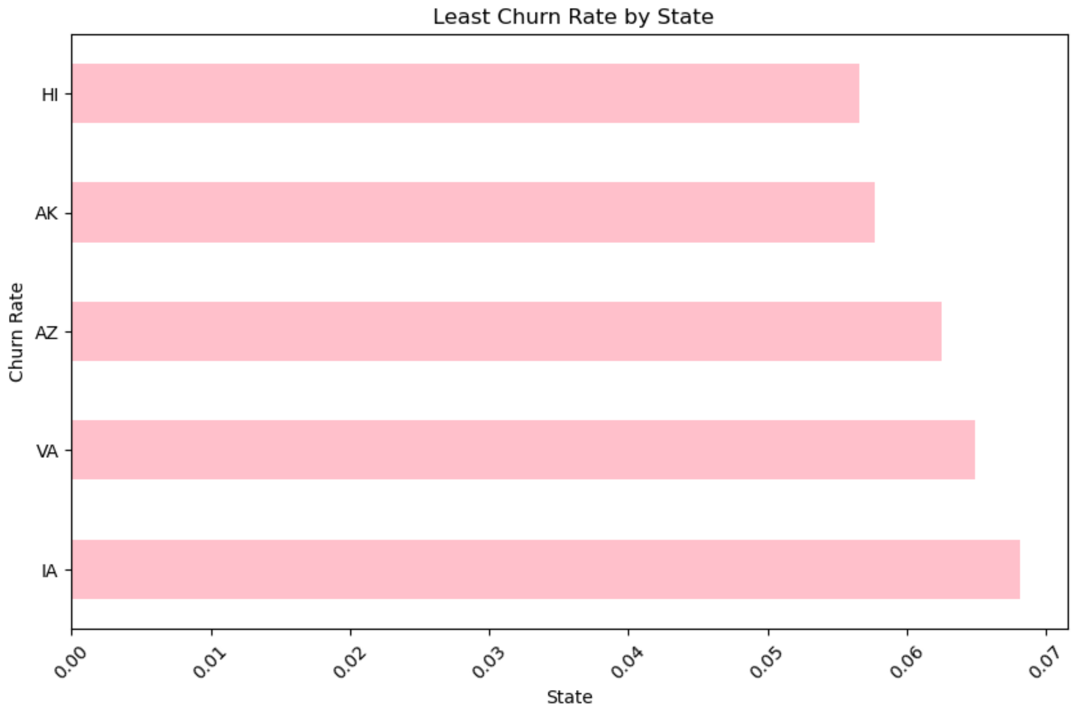
In **Texas (TX)**, Out of 72 customers, 18 churned this is a 25% churn rate.

In **Maryland (MD)**, Out of 70 customers, 17 churned which is a 24.3% churn rate.

In **South Carolina (SC)**, Out of 60 customers, 14 churned equal to 23.3% churn rate.

SyriaTel should prioritize churn reduction measures in these states (targeted offers, better customer service, tailored retention strategies).

2.3.3 States with the least churn rates



This bar plot shows the states with the least rates of churning. Below is their description:

In **Hawaii (HI)**, Out of 53 customers, only 3 churned equal to 5.7% churn rate.

In **Alaska (AK)**, Out of 52 customers, 3 churned which is 5.8% churn rate.

In **Arizona (AZ)**, Out of 64 customers, 4 churned adds upto 6.3% churn rate.

In **Virginia (VA)**, Out of 77 customers, 5 churned equals to 6.5% churn rate.

In **Iowa (IA)**, Out of 44 customers, 3 churned equals to 6.8% churn rate.

This is going to guide the company to know where its loyal customers are and where they have a stronger market.

Explanation

The baseline of churning rate was 14.49% so we used that to determine the churning limit to the features during the analysis above.

In the section above we were trying to identify the key factors that contribute to churning in the SyriaTel company and the states with the highest and lowest churning rates.

Before modelling we had a look at the columns and decided to work with the columns below:

where `churn` is our dependent variable. state – U.S. state where the customer lives. account length – Number of days the customer has had the account. area code – Telephone area code. phone number – Customer's phone number (serves as an identifier, not useful for prediction). international plan– Whether the customer has an international calling plan (yes/no). log_vmail_messages – Number of voicemail messages the customer has. customer service calls – Number of calls made to customer service.

churn – Whether the customer left the company (True = churned, False = stayed). which is our dependent variable total_calls - The total number of calls. total_minutes - The total number of minutes for all calls. total_charge - The total charges for all calls.

We merged the columns total day minutes, total eve minutes and total night minutes into one column named total_minutes. We also merged total day calls, total eve calls and total night calls into one column named total_calls. The columns total day charge, total eve charge, and total night charge were also merged to become one column called total_charge.

3. MODELING

In this section we were modeling models that would help SyriaTel predict the probability of a customer to churn or not to churn.

3.1 Logistic regression model

We chose our base model to be logistic regression simply because it is effective and efficient in predicting a binary classification of which our main task was a binary classification task.

Steps we took during modeling:

1. We first chose to use four variables with the highest positive correlation with our target variable.
2. We then defined our X as the four variables we chose and the y variable as our target variable.
3. Afterwards, we split the dataset into training and test set with a test size of 20%.
4. We then fitted our logistic regression model.

3.2 Logistic regression with all features

The next model we chose to use was a logistic regression model but with all the features.

We chose logistic regression again because it is effective and efficient in predicting a binary classification task.

We used the same steps as in our base model including all the features in the dataset.

3.3 Decision tree classifier

We wanted to capture the multicollinear relationship in our data that our logistic regression would have missed.

Steps we took:

1. We first chose to use all variables except for state.
2. We then defined our X as the variables we chose and the y variable as our target variable.
3. Afterwards, we split the dataset into training and test set with a test size of 30%.
4. In this step we initialized our model and performed a grid search before fitting our model.
    - We found our best parameters to be a cost complexity parameter of 0.082.
    - The criterion as entropy.
    - Max depth of 3.
    - Min samples leaf of 1.
    - Min sample split 2.

5. We then fit our model.

3.4 Random Forest classifier

We wanted to optimize our decision tree so we chose the random forest classifier.

Steps we took:

1. We first chose to use all variables.
2. We then defined our X as the variables we chose and the y variable as our target variable.
3. Afterwards, we split the dataset into training and test set with a test size of 30%.
4. In this step we initialized our model and performed a grid search before fitting our model.
   - We found our best parameters to be, estimators of 200.
   - Max depth of 5.
   - Min samples leaf of 2.
   - Min sample split 2.
5. We then fit our model.

4. EVALUATION

In this section we evaluated our models to determine which performs better at predicting churning customers. We compared the models and ultimately chose the one that performs better as our baseline model of recommendation.

We used Recall and ROC-AUC as the metric of success of our model. We used:

Recall

* Recall was used to measure how many actual churners the model correctly identifies.

* By optimizing high recall, we ensure the model captures most at-risk customers, even if it occasionally flags a few non-churners.

ROC-AUC

* Measures the model's ability to discriminate between churners and non-churners across all thresholds.

* ROC-AUC is threshold-independent, so it evaluates the model's overall ranking ability.

* A high ROC-AUC means the model is reliable in assigning higher churn probabilities to churners than to non-churners, which is critical for making informed business decisions.

Together they aligned with our business objectives and the problem we were trying to solve.

4.1 Logistic regression model evaluation

Starting with our base model we predicted our training and testing set and then evaluated our model using recall and ROC.

- Our findings were:
  1. recall:

  The model correctly identifies 75% of true churners meaning the model is good at catching churners even though it misses about 25%.
  2. It has an AUC of 81% which is good and shows a that our model is highly predictive but has room for growth.

4.2 Logistic regression with all features evaluation

This was our second model; we predicted our training and testing set and then evaluated our model using recall and ROC.

- Our model achieved a recall of: 71% and a ROC-AUC of: 80%

This model's metrics seem to have dropped as compared to our base model.

Our base model had a recall of 75% and a ROC of 81% while our second model has a recall of 71% and a ROC of 80%.

The model is slightly over-fitting: it performs better on training data than unseen test data. It means it will give biased insights and can't be used to make business insights. As seen from the accuracy below:

- Train Accuracy: 77.87%
- Test Accuracy: 74.96%
- Train ROC-AUC: 84.71%
- Test ROC-AUC: 80.39%

4.3 Decision tree classifier evaluation

This was our third model; we predicted our training and testing set and then evaluated our model using recall and ROC.

- The Decision Tree performed much better across all metrics than the baseline logistic regression.
- Recall: Slightly improved from 75.3% to 80.8% meaning the tree catches more actual churners.
- ROC-AUC: Higher AUC of 90% means the tree had a much better ability to discriminate churners from non-churners overall
- The Decision Tree clearly outperforms the baseline logistic regression and the second model on this dataset.

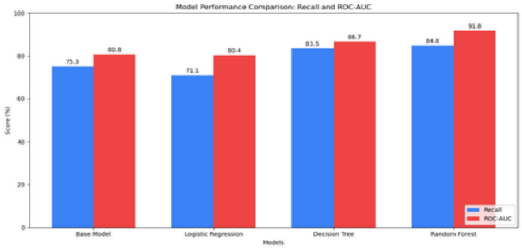
4.4 Random Forest classifier evaluation

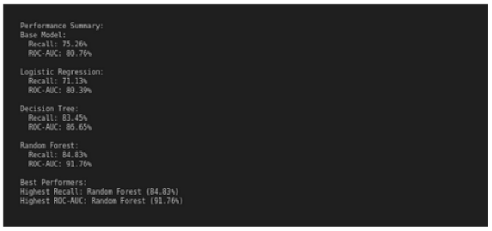
Random Forest outperforms the three models in recall, and ROC-AUC.

- ✓ It had a recall of 84.8%, but the precision and overall F1 improved significantly, meaning fewer false positives.

  ROC-AUC
- ✓ Random Forest has the highest ROC-AUC (91.7), indicating it discriminates churners from non-churners better than the other models.
- ✓ Random Forest gives the best overall performance on this dataset.

5. CONCLUSION





```
Performance Summary:
Base Model:
  Recall: 75.26%
  ROC-AUC: 80.76%

Logistic Regression:
  Recall: 71.13%
  ROC-AUC: 80.39%

Decision Tree:
  Recall: 83.45%
  ROC-AUC: 86.65%

Random Forest:
  Recall: 84.83%
  ROC-AUC: 91.76%

Best Performers:
Highest Recall: Random Forest (84.83%)
Highest ROC-AUC: Random Forest (91.76%)
```

1. The best performing model had a recall of 84.8% and an ROC-AUC score of 91.7%.
2. We chose the random Forest model because it achieved the best balance of a high recall (0.85) and ROC-AUC (0.92) with minimal over-fitting.
3. Random forest model predicted a total of 963 out 1000 correctly.

6. RECOMMENDATIONS

We would recommend our Random Forest model because it correctly identifies 84.8% of the actual churners. A high recall will ensure we catch more churners and this is important in churn prediction because missing churners (false negatives) can lead to lost revenue. ROC-AUC measures the model's ability to discriminate between churners and non-churners across all possible thresholds. Our score of 91.7% indicates that if you randomly pick a churner and a non-churner, the model assigns a higher probability of churn to the churner 91.7% of the time.

We can conclude that This model is very good at separating churners from non-churners compared to our baseline model and other models too.