SyriaTel Customer Churn

By Patience Severino

Business understanding

Customer retention is a problem for SyriaTel, a telecom provider. Customers that discontinue using a business's services are referred to as churn. Since keeping existing customers is frequently more economical than finding new ones, churn prediction is an essential task. The objective is to create a classifier that can forecast if a customer will "soon" cease doing business with SyriaTel.

Problem Statement.

Creating a predictive model that can precisely identify clients who are at danger of churning is the main challenge. The objective is to identify trends and indicators linked to churn by examining past customer data, such as service consumption patterns, invoicing details, and customer interactions. This knowledge will enable the company to apply focused retention tactics, like tailored promotions or enhanced customer service

Objectives

1. Develop a binary classification model that would predict whether a customer will churn or not.
2. Identifying features that would have an impact on customer churn.
3. Enhancing predictive accuracy by use of appropriate preprocessing and feature selection.
4. Optimizing model performance through feature engineering and tuning.
5. Determine churn prediction using metrics like accuracy and recall.

Success Criteria

1. Achieving a model accuracy and recall of at least 85% in churn.
2. Minimizing false positives to avoid unnecessary retention cost.

Modeling

The classification models that in order to predict customer churn on this dataset were;

- Logistic Regression
- Decision Tree
- Random Forest

The evaluation metrics that I focused on was recall since we are focused on customer retention and accuracy to determine how accurate the predictions are.

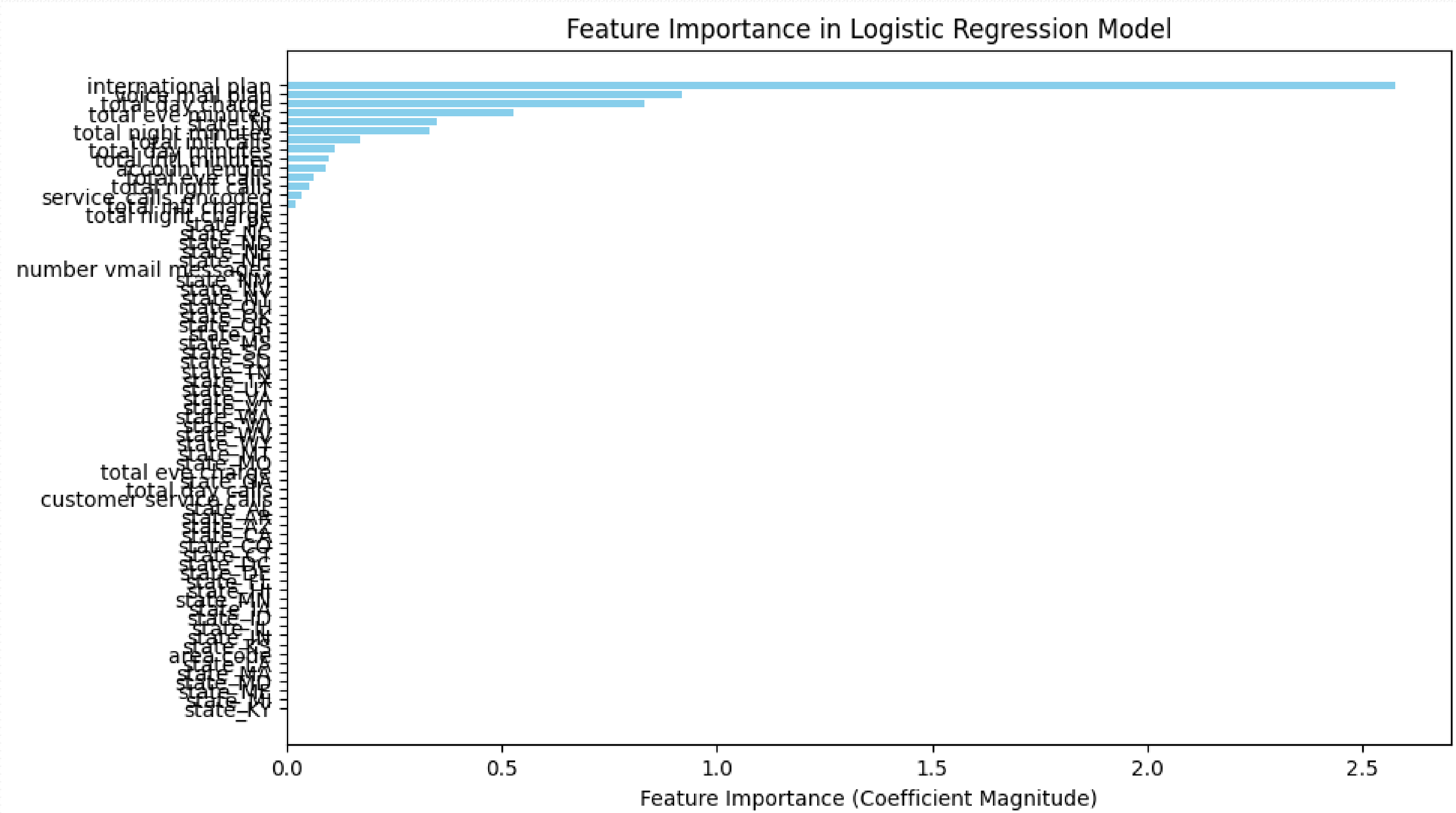
Logistic Regression

Logistic regression output;

The logistic regression function worked well with a recall of 97.98% and an accuracy of 89.9%.

The tuned logistic regression function worked well with an accuracy of 90.39% though recall reduced to 97.06%.the best parameters after tuning is {'C': 10, 'class_weight': 'balanced', 'penalty': 'l1', 'solver': 'liblinear'}.

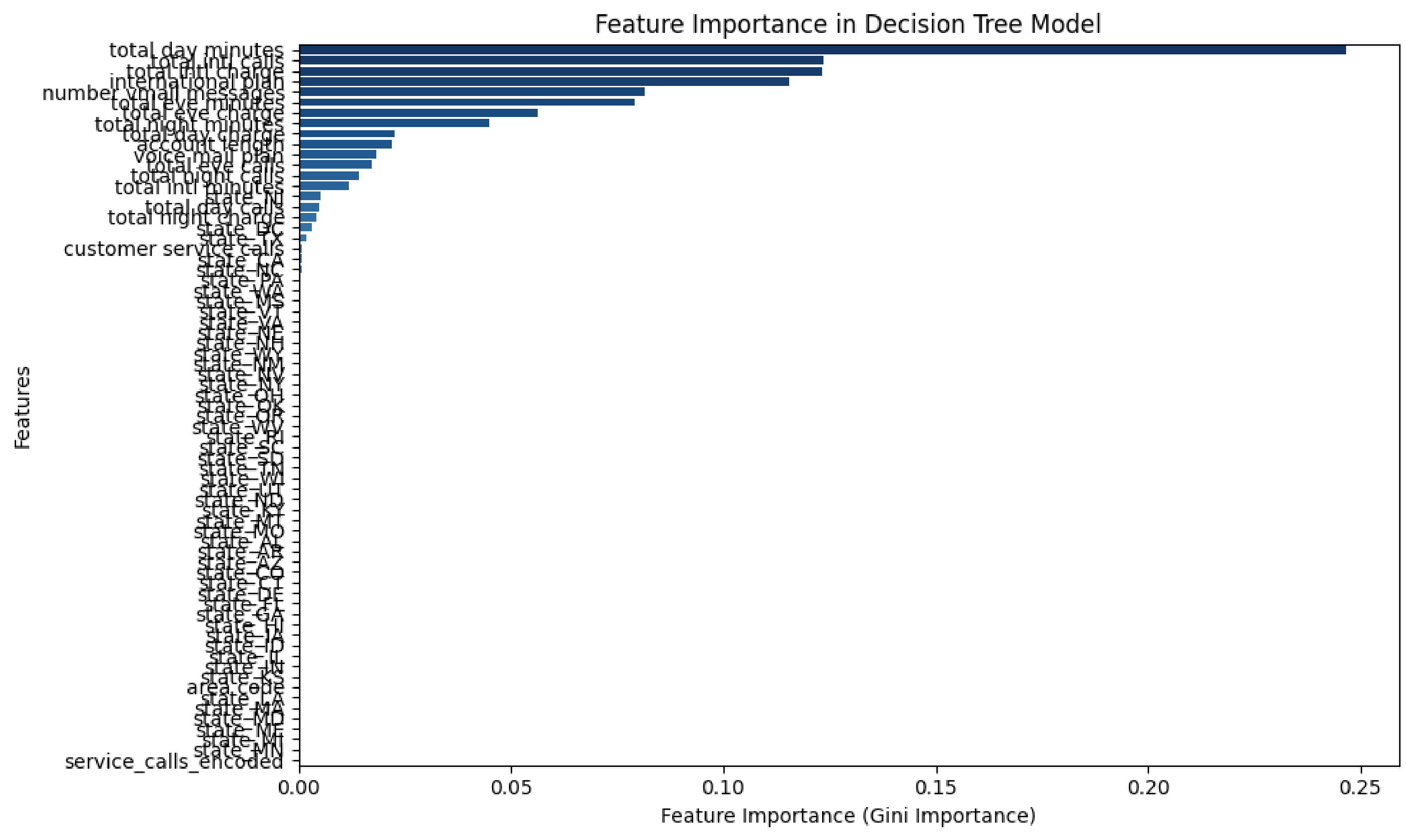
Feature importance in logistic regression



Decision tree

- Decision Tree output;
- The model worked well with an accuray of 92.01% and a recall of 95.04%.Let's check how the model works after tuning.Let's tune the model while adjusting its parameters to check if there is any improvement in the output.
- Tuned Decision Tree;
- The tuned model did better with an accuracy of 93.32% and Recall of 96.7%.The best parameters after tuning is {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}.

Feature importance in decision tree



Random Forest

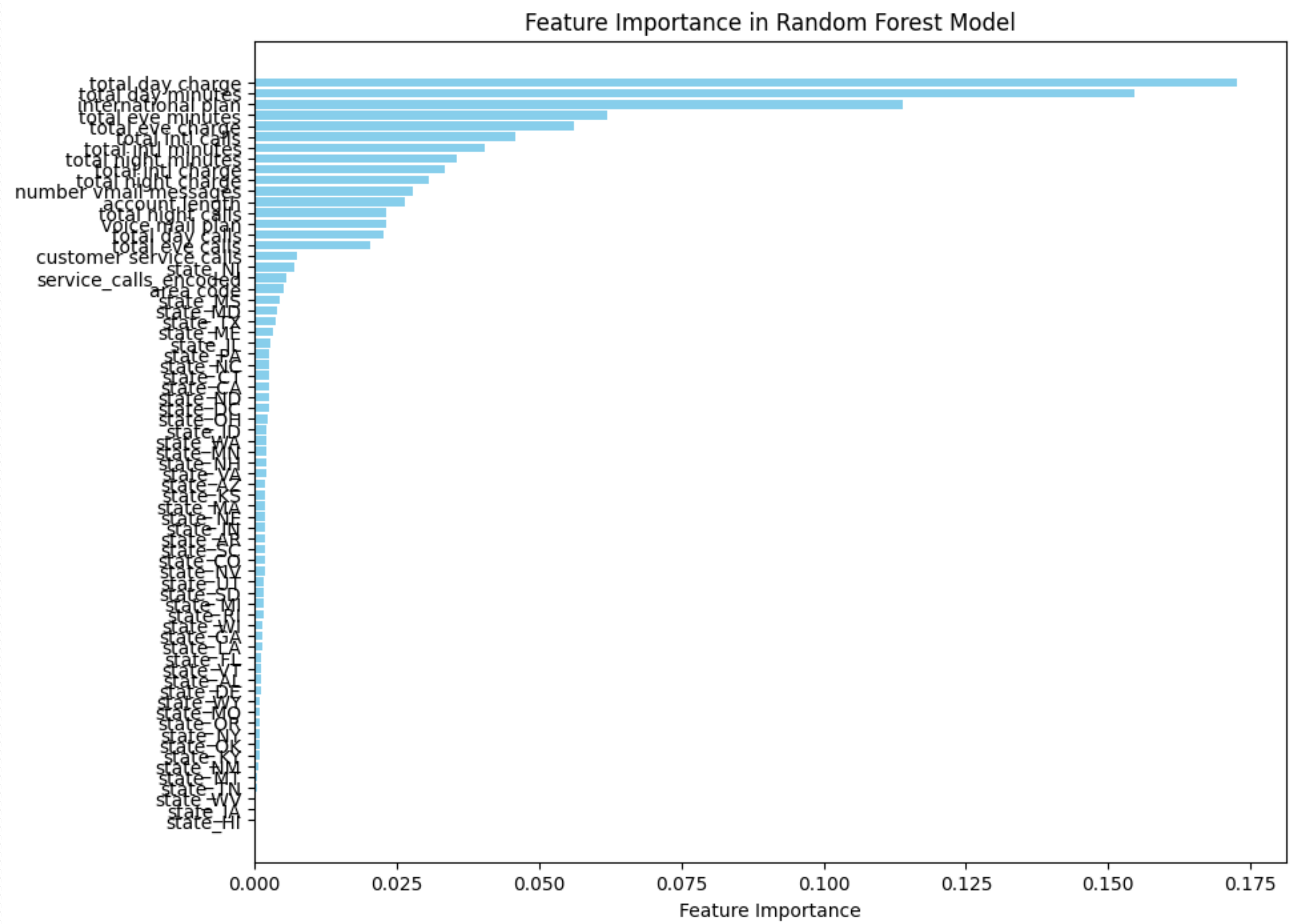
Random forest output;

The Random Forest function worked well with a recall of 99.63% and an accuracy of 94.79%.Let's tune the model while adjusting its parameters in order to see if there is an improvement in output.

The tuned random forest;

With the tuned model the accuracy reduced to 94.63% and the recall increased to 99.82%.The best parameters were 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 50

Feature importance in random forest



Model Evaluation

The metrics for evaluation that I focussed on with modelling is Accuracy and Recall.Recall is beneficial to the business since we are focusing on client retention.The best performing model is Random Forest with a Recall of 99.82% and accuracy of 94.63% Since we are focused on predicting churn and customer retention

Conclusion

Model Performance

The Random Forest model outperformed the other models, achieving a recall of 99.63% and an accuracy of 94.79% before tuning. After tuning, the recall remained strong, indicating its robustness in identifying churn cases accurately. The Decision Tree model also performed well, especially after tuning, with a recall of 96.7% and an accuracy of 93.32%.The Logistic Regression model, while offering insights into feature importance, showed slightly lower performance compared to tree-based models. The recall decreased to 97.06% after tuning, with an accuracy of 90.39%.

Feature Importance

The top features influencing churn varied across models: Logistic Regression: International plan, voicemail plan, and total day plan were significant. Decision Tree: Total day minutes, total international calls, and total international charge were prominent. Random Forest: Further validated the importance of these features, emphasizing total usage metrics and international service plans.

The major impact of the International Plan feature across all models indicates that customers with this plan are more likely to churn. This raises the possibility of a pricing or value perception conflict. Another important component of the Logistic Regression model was the Voicemail Plan. This plan may not be valuable to customers, which would increase churn.

Patterns of Call Usage

Two powerful indicators of churn are total international calls and total day minutes. High usage in these locations could be a sign that a better plan is needed or that call quality and pricing aren't meeting your needs.

Recommendations

1. Use the International Plan to develop exclusive deals or customer retention tactics. This could include better-value combined deals, discounts, or promotions. To better serve clients with high daytime usage and lower churn risk, create individualized strategies for them.

2. Strategies for Preventing Churn: For high-risk clients, put in place a proactive customer interaction program. When usage patterns resemble those of likely churners, for instance, offers may be sent or satisfaction surveys may be carried out

3. Deployment of the Model: For production use, implement the Random Forest model since it provides the optimal balance between recall and accuracy, guaranteeing a low number of false negatives in churn prediction.

4. Observation and Iteration: To maintain accuracy, keep an eye on the model's performance at all times and retrain it occasionally using new data. Create feature drift notifications, particularly if usage trends change over time.

Thank you!
patienceseverino36@gmail.com