Predicting Customer Churn in Telecommunications using Machine Learning

Deo Matchouli

I. INTRODUCTION & BACKGROUND

The world of business is more competitive than ever, each and every day it seems harder to keep a hold onto customer base & get products off the shelf, particularly in industries that form part of subscription model.

Churn: The rate at which customers cancel a given service or subscription. [4]

As it significantly affects the ability of businesses like telecom providers, streaming platforms & Software as a Service (SaaS) companies to maintain long term profitability & growth. Companies in these verticals spend extensively on customer acquisition and high early churn rates among those customers can wear away at a company's margins. Businesses need to have some idea as to who is likely to be leaving and what they can do pro actively rather than reactively. Predictive analytics [5] backed by machine learning models can help firms estimate customer turnover and act on it with precision.

II. RESEARCH QUESTION & SIGNIFICANCE

This project intends to use two machine learning models - logistic regression and random forest classification - to predict customer attrition. I will not only estimate whether a customer is likely to churn, but also identify the most influential elements contributing to churn behavior. By comparing the performance of these two models, I will see how different machine learning algorithms may be used to forecast churn.

Objectives

Accurate Prediction: I will create predictive algorithms to identify clients who are high risk of quitting the company.

Feature Importance: I will assess the major characteristics that influence customer churn, giving actionable insights for businesses.

By creating these models, I hope to provide organizations with a tool for early detection of prospective churners, allowing them to execute retention measures that can boost customer loyalty and reduce attrition.

Understanding and predicting client turnover is critical for businesses, particularly those that operate on subscription models. Here are few main reasons why firms need to accurately estimate churn.

Revenue Protection: Churn directly affects recurring revenue. Losing clients implies losing potential revenue streams. Companies that can forecast turnover can react proactively to avoid this loss.

Retention initiatives: Understanding turnover allows firms to create focused client retention initiatives. To keep high risk consumers engaged, tailor made promotions, discounts or even personalized services might be provided.

Operational Efficiency: Businesses can utilize limited resources more effectively by determining and avoiding the customers most likely to churn. Company could focus on retaining existing customers instead of investing heavily in new customer acquisition, which is often more cost effective.

The Competitive Advantage: Organizations must have the ability to sense early signs of customer dissatisfaction and do something about it. This could mean that they can keep more clients than their competitors by providing exceptional customer service and intervening at the right times.

III. LITERATURE REVIEW

To verse myself with the topic, I immersed myself in a wide range of studies and books on customer churn prediction, particularly in the telecommunications industry. Two key readings, Gerpott, T.J, Rams W & Schindler A. (2001). “Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market.” and Theodoridis, G. and Athanasios Tsadiras (2021). “Using Machine Learning Methods to Predict Subscriber Churn of a Web-Based Drug Information Platform.” were instrumental in deepening my understanding of the various models used in churn prediction. Research revealed [6] that while traditional models like logistic regression offer interpretability, machine learning models such as Random Forests and Gradient Boosting provide superior accuracy by capturing complex patterns. Additionally, the literature [7] highlighted challenges with class imbalance, often addressed by techniques like SMOTE and emphasized the importance of key features like customer tenure, service usage and billing information.[8] As big data technologies advance, studies also suggest that real time churn prediction is becoming increasingly feasible, offering businesses a more agile approach to customer retention.

IV. METHODOLOGY

A. Requirements

This study utilizes the Telecom Customer Churn dataset from Kaggle.[3] This dataset provides a full perspective of customer behavior, including the services to which each customer has subscribed, their demographic information, account details & churn rate.

Dataset Overview

Churn Indicator: The Churn column is the key target variable, indicating whether a client quit the service within the previous month.

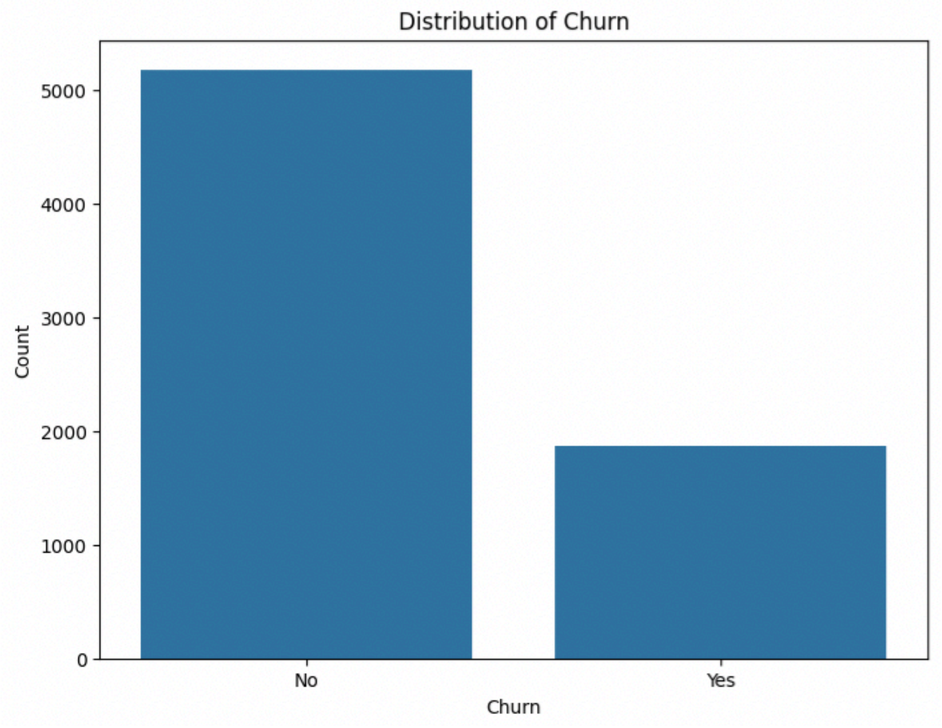
Services Subscribed: The dataset includes information about the services to which each client has subscribed, such as phone service, multiple lines, internet, online security, online backup, device protection, tech assistance and streaming TV & movies.

Account Details: Details on customer's account, including tenure (how long they've been with the company), contract type, payment methods, paperless billing options, monthly & total charges.

Customer Demographics: Dataset additionally contains demographic information such as gender, age, dependents or partners & senior status.

B. Explanatory Data Analysis

These visual aids offer preliminary understanding of the Telecom Customer Churn dataset by assisting in recognizing trends & connections that may impact customer attrition. These results can be the basis for more research & modeling to create successful predictive models.



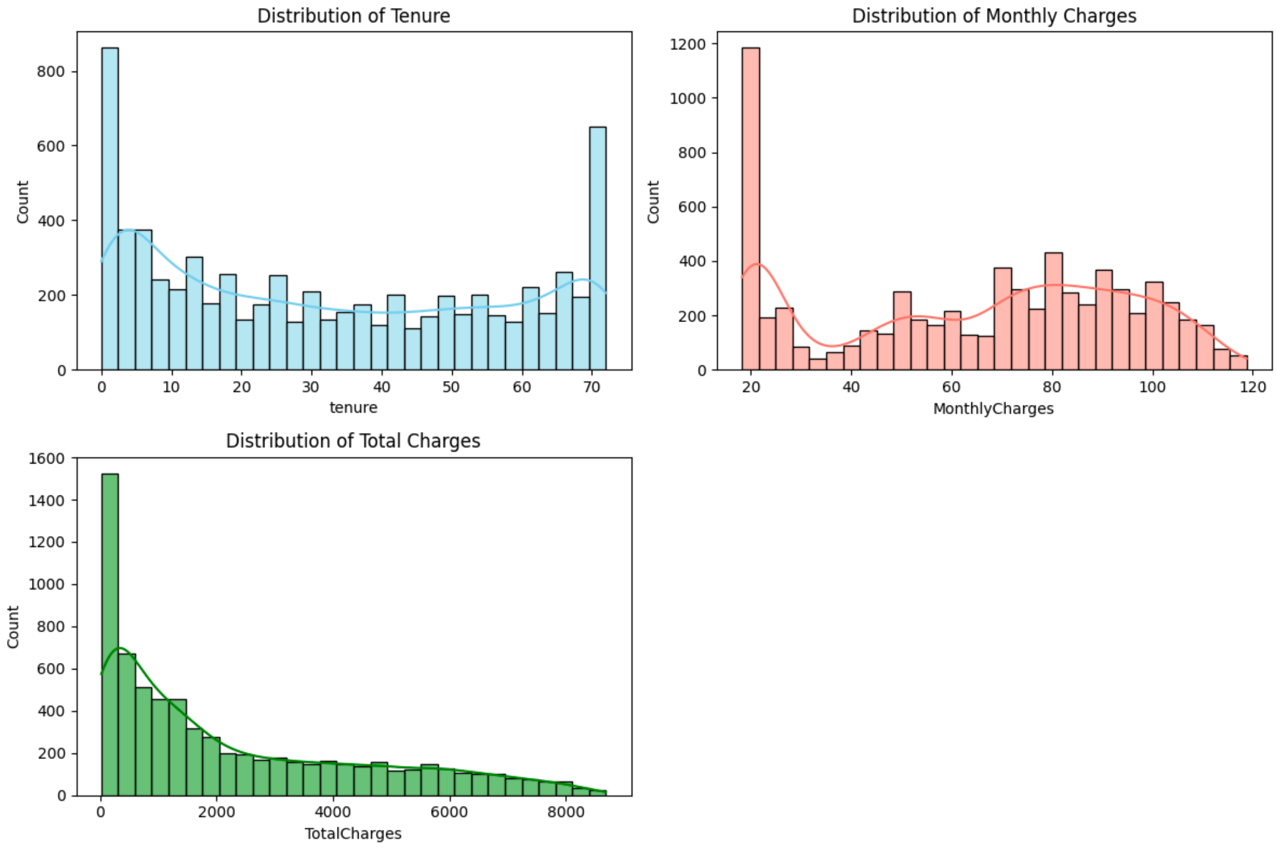
Analysis of churn

The threshold for No is substantially higher than the bar for Yes, indicating that the vast majority of customers do not churn.
Lower Rate of Churn: The lower height of the Yes bar indicates that the overall churn rate is low.

Several reasons may have contributed to the dataset's low churn rate,

High Customer Satisfaction: Customers who tend to be happy with the product or service are less likely to leave. Product quality, customer assistance and pricing can all have an impact on consumer happiness.
Strong Loyalty Programs: Loyalty programs that reward repeat consumers can encourage them to stick with the company.
Customers may be less likely to switch if there are only a few or no superior options accessible.
High transferring fees: Customers may experience significant fees or difficulties when transferring to a rival, which might reduce turnover.

Histograms representing the distributions of three key variables: Tenure, Monthly Charges & Total Charges.
These variables are likely relevant to customer churn in the telecommunications industry.



Tenure: Assists in determining the retention curve, which illustrates how a longer customer's duration reduces the chance of attrition.

A key component in comprehending churn is tenure. Shorter tenured customers may be more prone to leave because of disappointment or unfulfilled expectations.

Longer tenures are associated with decreased churn likelihood in logistic regression, where tenure and churn frequently exhibit a negative association. This is due to the fact that loyal clients have routines with the service and are typically more satisfied.

Plotting tenure versus churn probability makes it easier to see the point at which churn risk starts to decline noticeably. It offers insights into customer loyalty patterns.

The histogram shows a significant number of customers with relatively short tenures. This could indicate a high churn rate among new customers. However, there is also a substantial number of customers with longer tenures, suggesting that some customers remain with the company for extended periods.

Monthly Charges: Shows how sensitive consumers are to price, emphasizing key price points that have an impact on attrition.

The monthly fees have a direct impact on the financial hardship and customer happiness. If clients believe that greater fees are not worth the quality of the services they receive, this could result in higher churn rates.
Higher monthly charges are associated with a higher probability of churn, as there is often a positive correlation between them. This is due to the possibility of clients quitting if they believe they are being overcharged for the service.

A monthly fees plotted against the probability of churn shows how price fluctuations affect consumers' sensitivity. It assists in determining the price points where churn is most likely to occur.

The distribution of monthly charges is somewhat skewed to the right, indicating that a larger proportion of customers pay higher monthly fees. This could suggest that customers with higher charges may be more likely to churn due to cost considerations.

Total Charges: Illustrates the correlation between total spending and turnover, highlighting the effect of total money invested in the service.

The total charges show the customer's cumulative payment amount and thus the total financial commitment. Increased overall costs could be a sign of ongoing use and membership.
Total charges in logistic regression may exhibit a non-linear relationship with turnover. Customers may exhibit distinct churn behaviors depending on how much their total charges are. Because a higher overall fee represents a larger commitment in the service, it is generally associated with reduced turnover.
It is easier to observe how cumulative expenditure affects churn when total charges are plotted against churn probability. It draws attention to crucial moments when in response to financial commitment, the likelihood of churn varies considerably.

Making wise judgments to improve customer retention methods & comprehending the underlying trends in the data depend heavily on these representations. They enable us to identify particular client segments that are more likely to experience attrition & adjust treatments appropriately.

The histogram for total charges shows a similar pattern to monthly charges, with a rightward skew. This is likely due to the correlation between monthly charges and total charges over time.

Potential Implications for Churn

Based on observations, I can hypothesize the following factors may contribute to customer churn,

Short tenure: Customers who have been with the company for a shorter period may be more likely to churn due to dissatisfaction/lack of commitment.
High monthly charges: Customers who pay higher monthly fees may be more sensitive to price increases/seek more affordable alternatives.
Low perceived value: Customers who do not perceive the value of the service/feel that it does not meet their needs may be more likely to churn.

B. Design

Random Forest

An ensemble learning technique that builds multiple decision trees and averages their predictions. Random forest is robust to overfitting and can capture more complex relationships in the data.

Random Forest Classifier for predicting churn in this scenario is justified for several reasons based on its characteristics and performance

Ensemble Learning Capability: Random Forest is an ensemble technique that enhances generalizability and prediction accuracy by combining several decision trees.[9] When compared to individual decision trees, the forest is less likely to overfit since each tree functions independently.

Managing Categorical Variables: Without requiring a lot of preprocessing like one-hot encoding, Random Forests are capable of managing both numerical & categorical data. This is useful when working with datasets that contain categorical features.

Feature Importance Analysis: Random Forests offer a simple method for determining how significant various features are in predicting churn. This ability is essential for figuring out what aspects most affect consumer decisions and for developing focused retention and enhancement strategies.

Sturdy Performance: Random Forests often exhibit robust performance over a variety of datasets, including ones with missing or noisy data. Comparatively speaking, they are less susceptible to data outliers than other models like SVMs.

Scalability and Efficiency: Random Forests are usually scalable and can accommodate datasets with thousands of input variables without the need for variable deletion, despite the fact that they can be computationally expensive for large datasets.

Using a random forest classifier, we may rank the importance of each attribute. This is particularly beneficial for discovering non linear connections between variables and churn that logistic regression may not capture.

Logistic Regression

A linear model that assumes a direct relationship between the independent variables (features) and the dependent variable (churn). Logistic regression is interpretable and computationally efficient, making it a good baseline model. [10]

In terms of accuracy, precision, recall & ROC AUC, it typically outperforms Random Forest. This suggests that logistic regression would be a better model for this particular churn prediction task. Although logistic regression provides superior performance metrics other factors to take into account when selecting a model include computational efficiency, the interpretability of the results, the particular context and the necessary criteria.

This model allows us to analyze the coefficients of each feature, giving us a clear picture of the likelihood of churn based on many parameters.

V. IMPLEMENTATION/RESULTS

For implementation I used Colabatory notebook.

Random Forest

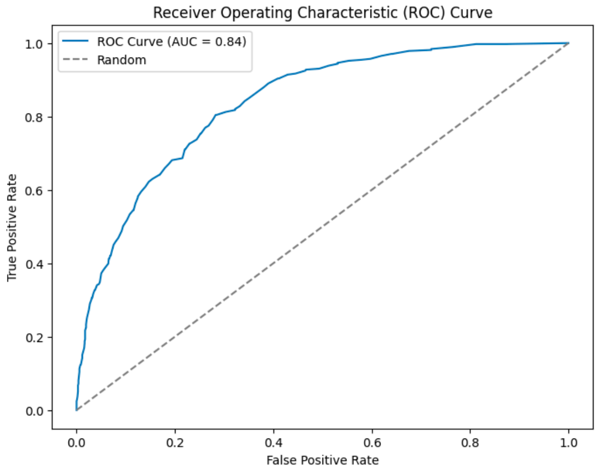
Random Forest Classifier metrics

Accuracy	**0.797**
Precision	**0.6654**
Recall	**0.4692**
ROC AUC	**0.8376**

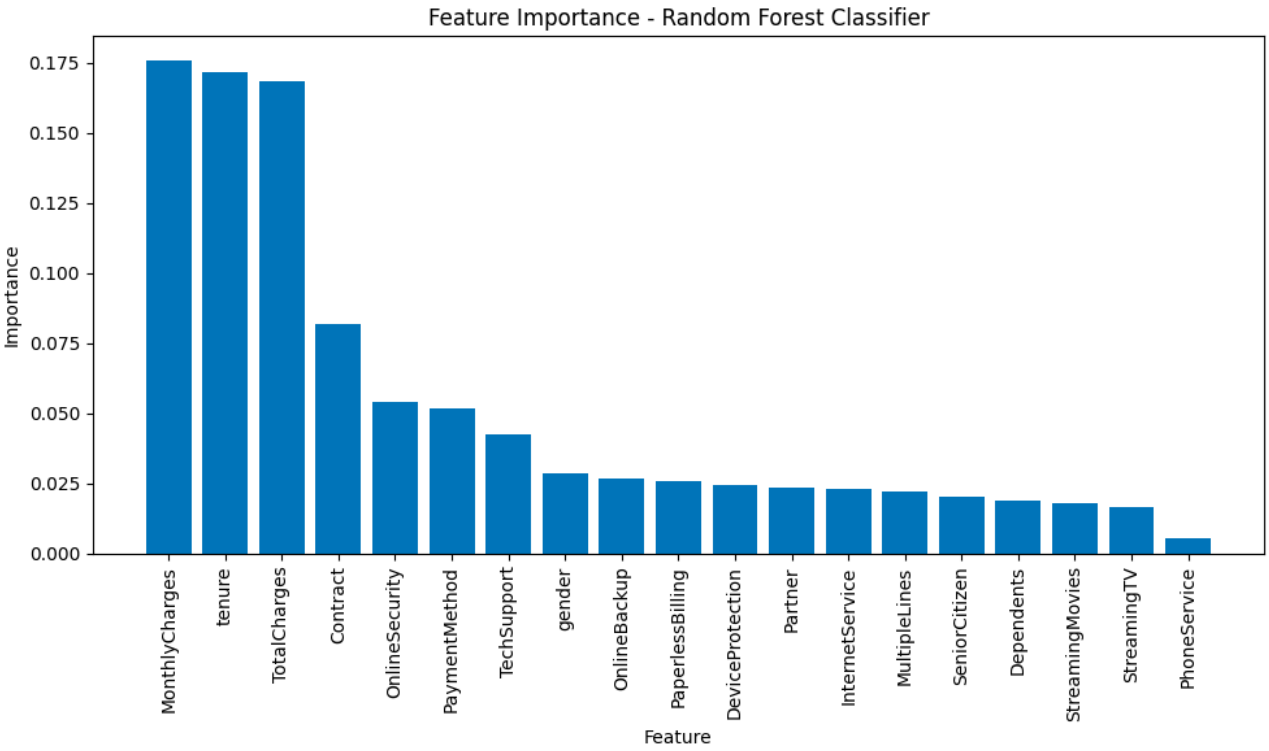
Analysis of the Results

Accuracy: This indicator measures the model's overall correct classification rate. In this situation, the accuracy is 79.7%, implying the model accurately predicted outcome for around 80% of data. Precision: This indicator represents the percentage of positive forecasts that were correct. A accuracy of 0.6654 suggests that 66.54% of the occurrences predicted by the model were actually positive.
Remember that this statistic counts the proportion of actual positive cases that were accurately predicted by the model. Recall of 0.4692 demonstrates that the model identified 46.92% of genuine positive cases.

ROC AUC: The ROC AUC (Area Under the Curve) is a measure of the model's ability to discriminate between positive & negative instances. ROC AUC of roughly 84% indicates the model has good discrimination power.



The ROC curve plots the true positive rate (sensitivity) against false positive rate (specificity) at different classification thresholds. A curve that is closer to the top left corner indicates better performance.
In this case, the ROC curve is relatively close to the top left corner, suggesting the model is able to effectively distinguish between positive and negative instances.



Key Observations

Monthly Charges: This feature is most important, suggesting monthly charges have a significant impact on customer churn.
Tenure: Tenure is also considered relatively important, indicating the length of time a customer has been with the company may be a factor in churn.
Total Charges: Total charges are also deemed important, which is likely due to the correlation with monthly charges.
Contract: The contract type appears to be a moderately important factor.
Online Security & Payment Method: These features also have moderate importance, suggesting that factors related to customer security and billing may influence churn.

The feature importance plot suggests that factors related to monthly charges, tenure, total charges and contract type are likely the most influential drivers of customer churn in this dataset.

To reduce customer attrition, strategies based on feature importance analysis can greatly enhance retention efforts. For high paying clientele and long term customers, personalized retention tactics, such as loyalty rewards and special care can strengthen their commitment. Clients can also be segmented by monthly fees, offering bundled services or premium options to add value & discourage switching providers.

Investments in improving service quality—especially online backup, tech support, & security—can boost satisfaction and reduce churn. Given the importance of internet services, plans should be optimized to align with customer needs. Offering flexible contract terms can accommodate various preferences, reducing dissatisfaction due to rigid agreements.

In addition, enhancing the convenience of payment methods (e.g. autopay) with incentives can improve customer satisfaction. Predictive analytics should be used to personalize communication based on customer demographics and service usage patterns (e.g. streaming habits) allowing for proactive churn prevention. Finally, customized advertising campaigns that target specific customer segments can increase engagement and retention, making marketing more effective. These insights, based on feature importance, can be used to create focused, data driven retention strategies.

Logistic Regression

Logistic Regression Metrics

Accuracy	**0.8155**
Precision	**0.684**
Recall	**0.563**
ROC AUC	**0.8601**

Accuracy: This metric represents overall correct classification rate of the model. In this case, the accuracy is 81.55%, which means model correctly predicted outcome for approximately 81.55% of data points.
Precision: This metric measures proportion of positive predictions that were actually correct. A precision of 0.684 indicates 68.40% of instances predicted as positive by model were indeed positive.
Recall: This metric measures proportion of actual positive instances that were correctly predicted by the model. Recall suggests the model was able to identify 56% of the true positive cases.
ROC AUC: ROC AUC is a measure of the model's ability to discriminate between positive and negative instances. An ROC AUC of 86% indicates that the model has good discrimination power.

VI. DISCUSSSION AND CONCLUSION

Evaluation

Metric	**Logistic Regression**	**Random Forest**
Accuracy	**0.8155**	**0.797**
Precision	**0.684**	**0.6654**
Recall	**0.563**	**0.4692**
ROC AUC	**0.8601**	**0.8376**

Accuracy

Logistic Regression has a slight edge in accuracy over Random Forest. This suggests that Logistic Regression made more correct predictions overall. However, the difference is not very large.

Precision

Logistic Regression also has higher precision than Random Forest. Precision measures how many of the predicted churns were actually true churns. This means Logistic Regression is better at reducing false positives (customers predicted to churn but who actually stayed).

Recall

Logistic Regression shows significantly better recall compared to Random Forest. Recall measures how many of the actual churns were correctly predicted. This indicates that Logistic Regression captures more of the customers who are truly at risk of churning, making it a better model for identifying at risk customers.

ROC AUC

Logistic Regression outperforms Random Forest in terms of ROC AUC. The ROC AUC score measures the model's ability to distinguish between churn and non churn cases across different thresholds. [11] A higher value indicates that the model is better at distinguishing between customers who will churn and those who will not.

Logistic Regression outperforms Random Forest in all key metrics accuracy, precision, recall and ROC AUC. The higher recall and ROC AUC scores make Logistic Regression particularly advantageous in identifying churn cases, which is crucial for retention strategies. Despite Random Forest being a more complex & a non linear model, its performance on this dataset is inferior, suggesting that the linear nature of Logistic Regression is sufficient for this problem.

Given these results, Logistic Regression is the preferred model for this churn prediction task, as it provides better predictive performance and interpretability. However, Random Forest could still be useful if further tuning or feature importance analysis is required to better understand the factors influencing churn.

VII. FUTURE WORK

For future work, I will explore advanced feature engineering, including interaction and temporal features, as well as testing ensemble models like stacking and algorithms such as XGBoost or LightGBM.
Addressing class imbalance with cost sensitive learning or anomaly detection can further enhance model performance.
Hyper parameter tuning can optimize model accuracy and responsiveness.
Incorporating deep learning models, interpretability tools like SHAP or LIME and cost sensitive modeling will provide both predictive power and actionable insights.
Lastly, A/B testing and behavioral segmentation could ensure my predictions lead to effective business interventions.

APPENDIX

https://www.kaggle.com/datasets/blastchar/telco-customer-churn - Dataset

https://github.com/dmch11/project/blob/main/Predicting%20Churn%20Using%20ML.ipynb - Github Repository

REFERENCES

[1] Gerpott, T.J., Rams, W. and Schindler, A. (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy, 25(4), pp. 249–269. doi:https://doi.org/10.1016/s0308-5961(00)00097-5.

[2] Theodoridis, G. and Athanasios Tsadiras (2021). Using Machine Learning Methods to Predict Subscriber Churn of a Web-Based Drug Information Platform. IFIP advances in information and communication technology, [online] pp.581–593. doi:https://doi.org/10.1007/978-3-030-79150-6_46.

[3] BlastChar (2017). Telco Customer Churn. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

[4] Roy, R. (2007). Computer assisted customer churn management: State-of-the-art and future trends. [online] Computers & Operations Research. Available at: https://www.academia.edu/1164313/Computer_assisted_customer_churn_management_State_of_the_art_and_future_trends.

[5] Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43(2), pp.204–211. doi:https://doi.org/10.1509/jmkr.43.2.204.

[6] Hugues makongote (2015). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. [online] Academia.edu. Available at: https://www.academia.edu/10812597/Churn_prediction_in_subscription_services_An_application_of_support_vector_machines_while_comparing_two_parameter_selection_techniques.

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16(16), pp.321–357. doi:https://doi.org/10.1613/jair.953.

[8] Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), pp.211–229. doi:https://doi.org/10.1016/j.ejor.2011.09.031.

[9] Breiman, L. (2001). Random Forests. Machine Learning, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

[10] Daniel, J. and Martin, J. (2023). Speech and Language Processing. [online] Available at: https://web.stanford.edu/~jurafsky/slp3/5.pdf.

[11] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, [online] 27(8), pp.861–874. doi:https://doi.org/10.1016/j.patrec.2005.10.010.