Prepared by Aditya, Sameer and Ashish

Customer Churn Prediction

Aditya Kumar Singh - RA2311003010916
Sameer Yadav - RA2311003010915
Ashish Kumar - RA2311003010905

Abstract

Problem:

The project's goal is to predict customer churn and identify the key factors driving it, empowering a company to proactively retain its most vulnerable clients.

Expected Outcome:

The expected outcome is a functional web application that accurately predicts the probability of a customer churning. This tool will provide key factors driving churn for individual customers

Method:

The methodology combines **survival analysis** to understand customer lifetime with a **Random Forest model** to predict churn. These insights are then made transparent with **explainable AI** and deployed via a Flask application.

Introduction

Background:

- Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the **cost of retaining an existing customer is far less than acquiring a new one.**

Motivation:

- The main motivation for this project is to provide a telecommunications company with a **powerful tool to combat customer attrition.** By predicting which customers are likely to churn. The project's ultimate goal is to **increase profitability by reducing customer turnover and maximizing the lifetime value of each customer**.

Importance:

- This project's importance is that it enables a proactive approach to customer retention, directly **boosting revenue and profitability by preventing churn**, which is more cost-effective than customer acquisition.

Existing System & Its Limitations

Current Methods:

- Survival Analysis: Used to model customer lifetime and the probability of churn over time, answering "when" and "why" customers are likely to leave.
- Machine Learning Models: A Random Forest model predicts if a customer will churn based on their characteristics and behavior, providing a clear "yes" or "no" answer.

Advantages of Manual/Traditional CAD:

- Proactive Strategy: Identifies at-risk customers early, allowing the company to retain them before they churn.
- Increased Profitability: By reducing customer loss, the company can boost revenue and profitability without the high costs of customer acquisition.

Limitations:

- Data Dependency: The model's accuracy is heavily dependent on the quality and completeness of the available historical customer data.
- External Factors: It cannot account for sudden, unforeseen events like a major service outage or new competitor pricing that could drastically change churn rates.

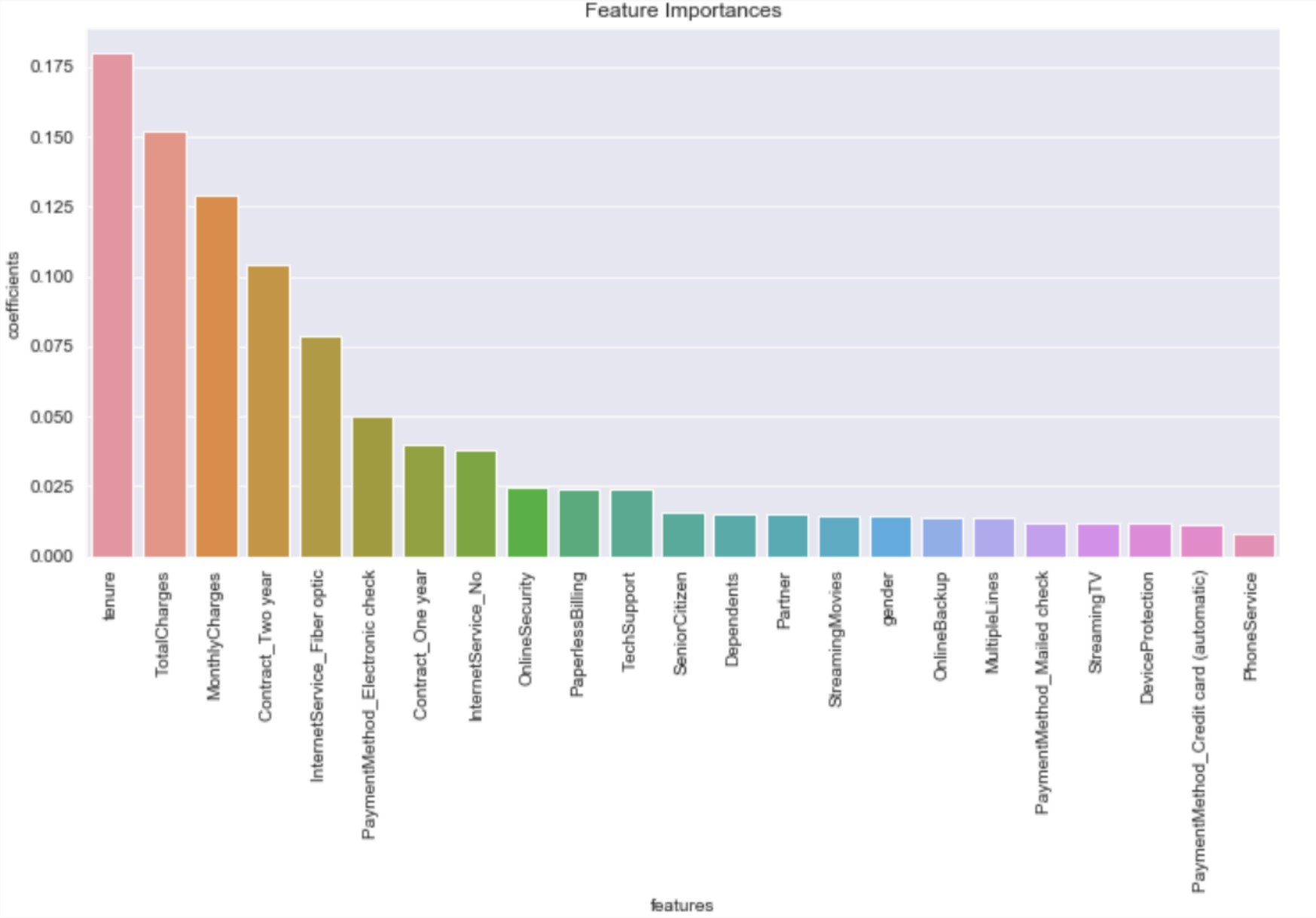
Dataset Used

- **Name: Telco Customer Churn Dataset**
- Number of Rows & Columns: The dataset contains information for **over 7,000 customers**. Specifically, it has **7,043 rows and 21 columns,** with each row representing a unique customer and each column containing a different attribute or feature.
- Data Types: The features in the dataset are a mix of different data types:
- Categorical Data: Most of the fields are categorical, like gender (Male/Female), Contract (Month-to-month, One year, Two year), and all of the service-related columns (e.g., OnlineSecurity, which can be Yes, No, or No Internet service).
- Demographics: Includes gender, SeniorCitizen, Partner, and Dependents.
- Numerical Data: There are a few key numerical columns, including tenure (number of months as a customer), MonthlyCharges, and TotalCharges.
- Customer Account Information: Includes customerID, tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, and TotalCharges.
- Target Variable: The Churn column, which indicates customer status.

Table

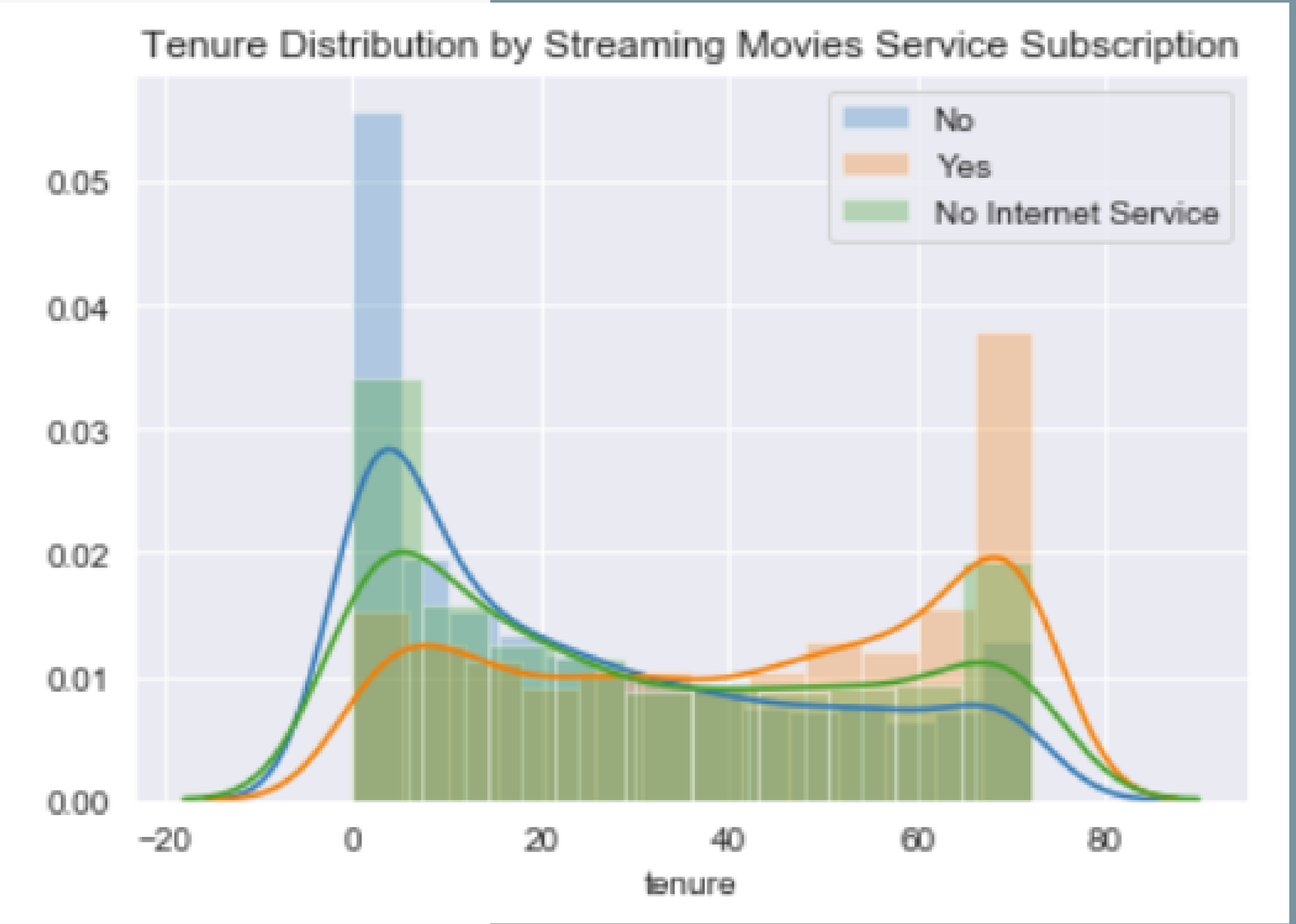
Customer ID	Gender	Tenure (Months)	Internet Service	Monthly Charges ($)	Churn
7590-VHVEG	Female	1	DSL	29.85	No
5575-GNVDE	Male	34	DSL	56.95	No
3668-QPYBK	Male	2	DSL	53.85	Yes
7795-CFOCW	Female	45	DSL	42.30	No
9237-RQOAY	Male	10	Fiber optic	98.75	Yes
9305-CDWPC	Female	35	Fiber optic	105.00	No
5917-WSDOQ	Male	4	DSL	45.10	Yes
2748-UQRNN	Male	13	DSL	70.30	No
7892-POXCK	Female	68	Fiber optic	109.90	No

Statistical Analysis - Data Overview



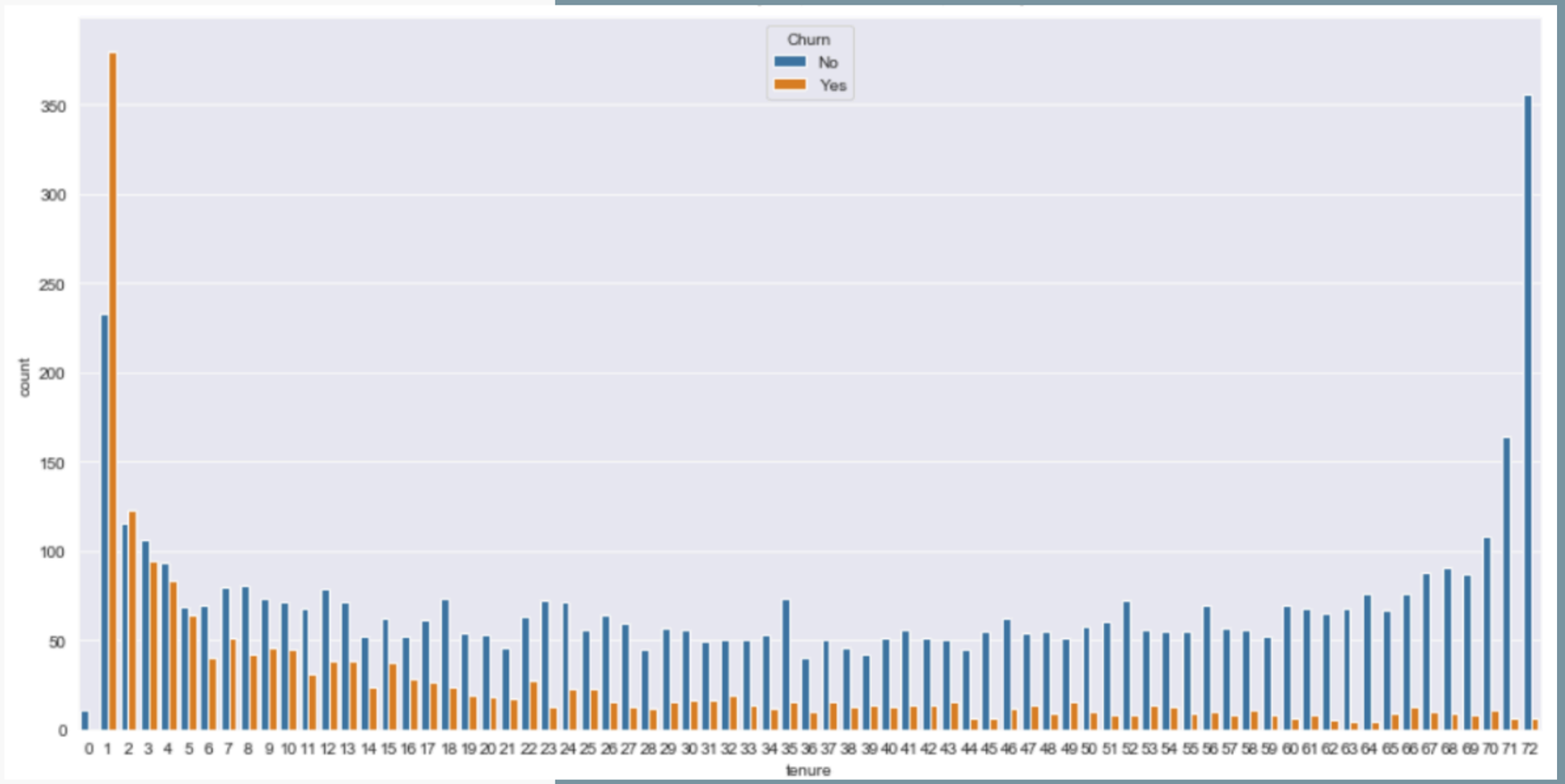
In this feature importance plot, we can see which features govern the customer churn.

Statistical Analysis - Data Overview



When the customers are new they do not opt for various services and their churning rate is very high. This can be seen in above plot.

Statistical Analysis - Data Overview



Churn and Tenure Relationship:

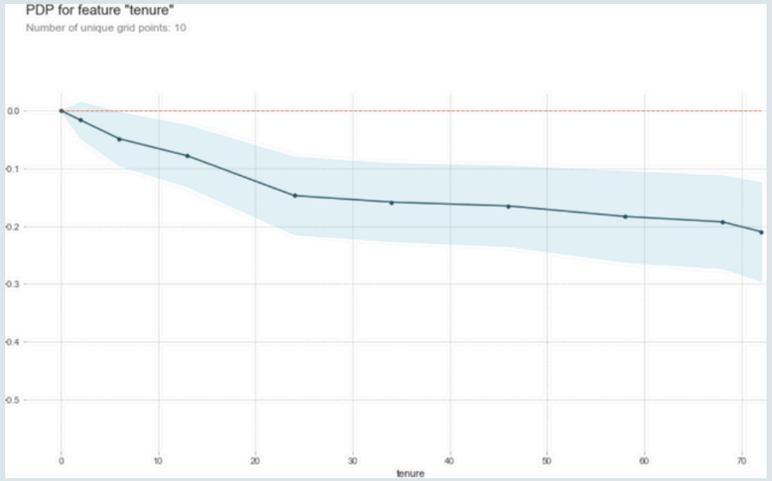
As we can see the higher the tenure, the lesser the churn rate. This tells us that the customer becomes loyal with the tenure.

Explainable AI modules using Random forest models-

Permutation Importance

Weight	Feature
0.0185 ± 0.0058	InternetService_Fiber optic
0.0064 ± 0.0088	Contract_Two year
0.0045 ± 0.0058	OnlineSecurity
0.0041 ± 0.0134	Contract_One year
0.0038 ± 0.0086	PaymentMethod_Electronic check
0.0037 ± 0.0071	InternetService_No
0.0028 ± 0.0094	tenure
0.0026 ± 0.0011	OnlineBackup
0.0020 ± 0.0078	MonthlyCharges
0.0010 ± 0.0014	DeviceProtection
0.0009 ± 0.0083	PaperlessBilling
0.0007 ± 0.0030	TechSupport
0.0004 ± 0.0032	StreamingMovies
0.0003 ± 0.0017	gender
0.0001 ± 0.0019	PhoneService

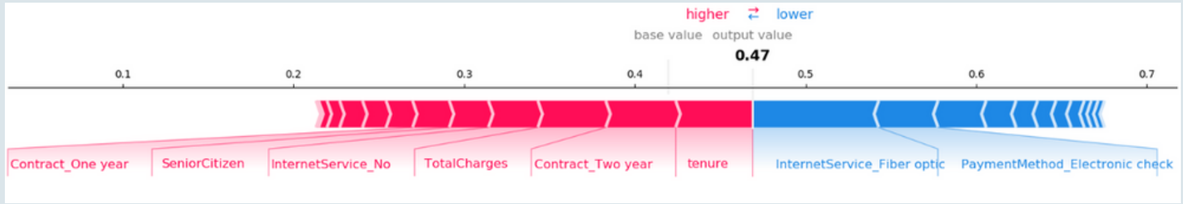
Permutation Importance shows feature importance by randomly shuffling feature values and measuring how much it degrades our performance.



Partial Dependence plots

Partial dependence plot is used to see how churning probability changes across the range of particular feature. For example, in side graph of tenure group, the churn probability decreases at a higher rate if a person is in tenure group 2 compared to 1.

Shap values



Shap values (Shapley Additive explanations) is a game theoretic approach to explain the output of any machine learning model. In above plot we can see that why a particual customer's churning probability is less than baseline value and which features are causing them.

Conclusion and Future Work

Conclusion:

- This project delivers an end-to-end customer churn prediction solution, combining survival analysis, machine learning, and explainable AI to provide predictive models and actionable insights.

Result:

- The project is expected to significantly reduce customer churn, leading to a direct increase in revenue and profitability by allowing proactive and targeted customer retention efforts.

Future Plans:

- Model Retraining: Regularly retrain the model with fresh customer data to maintain its accuracy and adapt to changing market dynamics.
- Advanced Analytics: Integrate more complex features, such as customer support interaction history or call detail records, to further improve the model's predictive power.

References

- **Telco Customer Churn Dataset** (Kaggle, originally from IBM) used for analysis and predictive modeling.
- Lundberg & Lee (2017): Introduced SHAP, the **explainable AI** method for interpreting the Random Forest model.
- Breiman (2001): Original paper on **Random Forest algorithm**, used for churn prediction.
- Cox (1972): Foundational work on Cox-proportional Hazard model, applied in **survival regression analysis.**

Thank you