EXPLAINABLE AI (XAI) FOR ENHANCED TRANSPARENCY AND TRUST IN MACHINE LEARNING MODELS

S1-24_SEZG628T: Dissertation

by

Harshit Jindal

(2022MT93524)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

VIDYA VIHAR, PILANI, RAJASTHAN – 333031

November 2024

EXPLAINABLE AI (XAI) FOR ENHANCED TRANSPARENCY AND TRUST IN MACHINE LEARNING MODELS

S1-24_SEZG628T: Dissertation

by

Harshit Jindal

(2022MT93524)

Submitted in partial fulfilment of the requirements for the degree of

M.Tech. Software Engineering

Under the supervision of

Mr. Mohit Agarwal

American Express

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

VIDYA VIHAR, PILANI, RAJASTHAN – 333031

November 2024

CERTIFICATE

This is to certify that the Dissertation entitled “Explainable AI (XAI) for Enhanced Transparency and Trust in Machine Learning Models” submitted by Harshit Jindal holding BITS ID No. 2022MT93524 for the partial fulfillment of the requirements for the degree of M.Tech. Software Engineering at BITS Pilani, embodies the bonafide work done by him under my supervision.

_____

Signature of the supervisor

Mohit Agarwal
Sr. Manager – Digital Product Management
American Express
Date: November 18, 2024

ABSTRACT

Explainable Artificial Intelligence (XAI) has emerged as the need of the hour to ensure transparency and trust in machine learning models, including within the financial sector where decision-making processes must be both accurate and interpretable. This study explores various XAI methodologies applied to predictive models using the Bank Churn Dataset from Kaggle. By implementing inherently explainable models such as Logistic Regression, Decision Tree and Explainable Boosting Machine (EBM), alongside complex models like Random Forest, and applying post-hoc explainability techniques including LIME and SHAP, this research evaluates the balance between model performance and interpretability. Utilizing Scikit-learn and Microsoft's InterpretML platform, the study touches upon how different XAI tools enhance understanding of model predictions, meeting both regulatory requirements and business needs. While complex models may offer higher accuracy, the study seeks to assess if explainable models and advanced XAI techniques provide sufficient transparency without substantial loss in performance. This balance is crucial for financial institutions aiming to leverage AI responsibly, ensuring compliance, and maintaining stakeholder trust.

Keywords: Explainable AI (XAI), Machine Learning, Interpretability, Explainability, Regression Models, LIME, SHAP, Microsoft InterpretML, Explainable Boosting Machine (EBM).

_____	_____
Signature of the student	Signature of the supervisor
Harshit Jindal	Mohit Agarwal

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has guided and supported me throughout the journey of working on my master's dissertation on "Explainable AI (XAI) for Enhanced Transparency and Trust in Machine Learning Models."

To begin, I extend my sincerest thanks to Priyesh Jain, who is my additional examiner for this dissertation. His insightful feedback and constructive suggestions have been invaluable in refining and enhancing the overall quality of this work.

I am also profoundly grateful to my supervisor, Mohit Agarwal, for his unwavering support, encouragement, and thoughtful guidance. His advice and continuous support have helped steer this project forward, and I deeply appreciate the time and effort he dedicated to overseeing my work.

I would like to acknowledge Prof. Sanjay K. Sahay, my esteemed guide from BITS Pilani – Goa Campus, for the mentorship and academic foundation he provided. His advice and expertise have been crucial in navigating the complexities of this research.

Finally, I appreciate all the faculty members and staff at BITS Pilani for providing the necessary resources and a conducive environment for research and learning.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

1. AI – Artificial Intelligence
2. AUC – Area Under the Curve
3. EBM – Explainable Boosting Machine
4. F1 – F1 Score (A metric combining precision and recall)
5. GDPR – General Data Protection Regulation
6. ID3 – Iterative Dichotomiser 3 (Decision Tree Algorithm)
7. LIME – Local Interpretable Model-agnostic Explanations
8. ML – Machine Learning
9. SENN – Self-Explainable Neural Network
10. SHAP – Shapley Additive exPlanations
11. XAI – Explainable Artificial Intelligence
12. SMOTE – Synthetic Minority Over-sampling Technique

CHAPTER 1: INTRODUCTION

1.1 Background and Motivation

Over the past decade, the world has witnessed a massive surge in the adoption of Machine Learning for various use cases spanning almost all industries, with some like the financial sectors being at the forefront of this transformation. Most financial organization today use ML in some shape to power their processes, whether it be for credit decisioning, fraud detection, or algorithmic trading. However, there are also very stringent regulatory frameworks that mandate transparency in this decision-making process, and businesses must adhere to these in addition to focusing on model performance and accuracy. This is for a good reason, especially when these decisions can significantly impact consumers, such as loan approvals and credit assessments.

1.2 Problem Statement

While modern ML modelling techniques offer remarkable prediction accuracy, the more complex they get, the less interpretable they become. It is clear why this could potentially hinder their widespread adoption in critical financial applications, and thus the challenge lies in balancing model complexity with interpretability to meet regulatory requirements and maintain stakeholder trust without compromising performance.

1.3 Objectives of the Study

1. Explore the different XAI techniques to aid in ML model explainability.
2. Implement both inherently explainable models, as well more complex black-box models and apply XAI techniques on them.
3. Provide actionable insights that businesses can adopt to reap the maximum benefits that ML has to offer, while also maintaining transparency in their processes.
4. Assess the tradeoff between model complexity and its interpretability.

1.4 Scope and Limitations

1. The study explores both inherently explainable models, as well as complex black-box modeling algorithms and the application of XAI techniques to make them explainable.
2. The research employs the open-sourced “Credit Card Customers” dataset (with a CC0: Public Domain License) from Kaggle to predict churning customers, providing a suitable context for evaluation of the effectiveness of various XAI techniques in predicting customer behavior within the financial sector.
3. The thesis assesses both model performance (using metrics like Accuracy, F1 Score, Precision, Recall) and explainability of the model predictions using visuals, data points, and feature importance.
4. The study may not capture the full complexity and diversity of real-world financial data since it is limited to the Kaggle dataset.

1.5 Dissertation Organization

Chapter 1: Introduction – Introduces the research context, problem statement, objectives, and significance.

Chapter 2: Literature Review – Reviews existing literature on XAI and its applications.

Chapter 3: Research Methodology – Outlines the research design including the models, techniques, and performance metrics to be used.

Chapter 4: Dataset Exploration and Preprocessing – Discusses the balance between model performance and explainability.

Chapter 5: Experimental Model Evaluation – Presents and analyzes the findings from implementing various models and XAI methods.

Chapter 6: Conclusion and Discussion – Summarizes the study and interprets the results in the context of research questions and objectives.

References – Lists all sources cited in the thesis.

Appendices – Provides supplementary material supporting the research and development work.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to Explainable AI (XAI)

With a surge in the use of AI for solving increasingly complex problems such as Large Language Models (LLMs), Medical Imaging, and Financial Decisioning, it is prudent to come up with ways to be able to understand what goes on inside a model. It is not wise to treat it as a mere black box of decision making when real human lives depend on it. Explainable AI (XAI) tackles this problem exactly. It refers to a set of processes and methods that can be leveraged to make AI models more transparent so we can answer questions such as how the model makes an output, and why. This understanding is crucial when it comes to analyzing model working, and build trust that we are legally and ethically doing the right thing.

2.2 Types of Explanations

1. Local vs. Global Explanations
    a. Local Explanations focus in on individual predictions or what is called an instance of a model. These help us understand the factors and reasons behind a particular prediction being made, and can extend to what changes would have resulted in a different outcome [3][5].
    b. Global Explanations offer a high-level view of the model's behavior. These can highlight general patterns in the model's outputs and show which features are more

important than the others without referring to a particular prediction [3][5].

2.3 Broad Approaches to XAI

1. Ante-hoc vs. Post-hoc Methods
    a. Ante-hoc Methods are concerned with building models with interpretability in mind. These models are inherently transparent, and the decisions made by them are explainable without sophisticated techniques.
        i. Logistic Regression: This technique involves fitting an equation to the training data by tweaking the "weights" associated with each feature. By just looking these weights, we can understand a lot about the model such as which attributes are more important than others and how much the resultant score will change by introducing a change in particular attribute [1][6].
        ii. Decision Trees: These refer to a flowchart-like model consisting of nodes and edges. Each node is a decision based on the value of an attribute, and accordingly, the flow can move along the edges of the node. The leaf nodes represent the decision of the model, hence making it clear how a prediction was made [1][6].
        iii. Explainable Boosting Machines (EBMs): Invented by Microsoft, this is a method of implementing gradient boosting algorithms that are inherently explainable. These are meant to achieve the performance generally associated with more complex models, while still retaining the interpretability aspect of simpler models [1][6].

b. Post-hoc Methods: Post-hoc methods come into the picture after a model is trained, especially when dealing with complex and non-interpretable modeling techniques.
    i. LIME (Local Interpretable Model-agnostic Explanations): LIME explains individual predictions by approximating the complex model locally with a simpler and more interpretable one, which is called a surrogate model. It provides insights into why a specific decision was made, one instance at a time, solving for the local interpretability needs [4][8].
    ii. SHAP (SHapley Additive exPlanations): Borrowing from game theory, SHAP assigns each feature an importance value based on its contribution to the prediction, and can help with the global explainability aspect of the model. SHAP ensures consistency and fairness in feature attribution [4][8].

2. Model-Agnostic vs. Model-Specific Methods
    a. Model-Agnostic Methods, as the name suggests, can be applied to any model. They treat the model as a black box and focus on the input-output relationship rather than the model architecture or the techniques used to train it. LIME and SHAP are prime examples of this method [4][7].
    b. Model-Specific Methods are tailored to particular algorithms, these methods leverage the internal structure of the model for explanations. For instance, feature importance in random forests or attention weights in neural networks.

2.4 Regulatory Context

Authorities worldwide are calling for an increased scrutiny and regulations on AI models to safeguard the interests of humanity. They demand transparency to protect consumers and ensure fairness in the usage of models by large enterprises to make decisions.

1. General Data Protection Regulation (GDPR): Enacted by the European Union, GDPR includes a "right to explanation," meaning individuals can ask for explanations of automated decisions affecting them [2][6].
2. Fair Credit Reporting Act (FCRA): In the context of financial regulations, the FCRA in the United States promotes accuracy, fairness, and privacy in credit reporting. It indirectly pushes for explainable models in credit-related decisioning [2][6].

2.5 Challenges and Gaps

Despite the progress, several hurdles impede the full realization of XAI's potential.

1. Complexity vs. Interpretability: There's often a trade-off between a model's performance and its transparency. Complex models like deep neural networks offer higher accuracy on more complex problems are come with their own set of challenges in terms of interpretability [1][5][7]. Hence, it is generally preferred to use simpler models if and when they perform equally well, and this concept is referred to as the Occam's Razor [5].
2. User Diversity: Different stakeholders have varying needs. What satisfies a data scientist might baffle a customer. Creating explanations that cater to diverse audiences remains a challenging problem [2][10].

3. Scalability Issues: Techniques that work on small datasets may not scale very well in an enterprise setting. For example, financial institutions handle massive amounts of data, necessitating scalable XAI solutions [7][10].

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Research Design

This study focuses on quantitative research along with exploratory and comparative analysis with a goal to evaluate the effectiveness of different XAI techniques in enhancing transparency and trust in machine learning models, including but not limited to the financial domain.

1. Descriptive Analysis: To understand the characteristics and distribution of the Kaggle's "Credit Card Customers" dataset to predict customer attrition.
2. Predictive Modeling: Utilizing different machine learning algorithms to predict customer churn.
3. Explainability Assessment: Applying ante-hoc and post-hoc XAI techniques to interpret model predictions.
4. Comparative Analysis: Evaluating the balance between model performance and interpretability across different models and XAI methods.

3.2 Models and Techniques

Once the dataset has been analyzed, cleaned, split into training and testing partitions, and ready for model training, the following models are trained on it in increasing order of complexity:

1. Inherently Explainable:
    a. Logistic Regression

    b. Decision Tree
    c. Explainable Boosting Machine (EBM)
2. Black Box:
    a. Random Forest

The inherently explainable models are built using Scikit-learn and Microsoft's InterpretML libraries, and the local and global interpretations are plotted to understand the feature importance for each of them. Random Forest is implemented as a complex black-box model on which LIME and SHAP are applied to analyze the predictions. For each model type, the confusion matrix, accuracy, precision, recall and f1-scores act as the metrics for capturing and comparing the model performance. Finally, the results are compared to draw and discuss the conclusions.

CHAPTER 4: DATASET EXPLORATION AND PREPROCESSING

4.1 Dataset Description

The open-sourced dataset used in this study is available on Kaggle under the CC0: Public Domain license (Appendix). It contains 10,000 records with 23 columns, out of which 20 are independent features, 1 is the output flag to represent customer attrition, and 2 are extra features that need to be dropped. However, only 16.07% customers have churned making the model training challenging.

4.2 Dataset Attributes

The following attributes are available in the dataset:

Table 1: Attributes for "Credit Card Customers" Dataset from Kaggle

Sr. No.	Attribute	Datatype	Description
1.	clientnum	Numeric	Unique identifier for each customer
2.	attrition_flag	Categorical	Output flag indicating whether the customer is an “Existing Customer” or an “Attrited Customer”
3.	customer_age	Numeric	Age of the customer

4.	gender	Categorical	Gender of the customer (“M” for male, “F” for female)
5.	dependent_count	Numeric	Number of dependents the customer has
6.	education_level	Categorical	Highest level of education completed by the customer (e.g., “High School”, “Graduate”, “Uneducated”)
7.	marital_status	Categorical	Marital status of the customer (e.g., “married”, “single”)
8.	income_category	Categorical	Customer’s income category (e.g., “$60K - $80K”, “Less than $40K”)
9.	card_category	Categorical	Type of credit card the customer holds (e.g., “blue”, “gold”, “silver”)
10.	months_on_book	Numeric	Number of months the customer has been with the bank
11.	total_relationship_count	Numeric	Total number of products held by the customer with the bank
12.	months_inactive_12_mon	Numeric	Number of months the customer has been inactive in the last 12 months
13.	contacts_count_12_mon	Numeric	Number of contacts with the customer in

			the last 12 months
14.	credit_limit	Numeric	Credit limit of the customer
15.	total_revolving_bal	Numeric	Total revolving balance on the customer's credit card
16.	avg_open_to_buy	Numeric	Average available credit for the customer over last 12 months
17.	total_amt_chng_q4_q1	Numeric	Change in transaction amount between Q4 and Q1
18.	total_trans_amt	Numeric	Total transaction amount in the last 12 months
19.	total_trans_ct	Numeric	Total number of transactions in the last 12 months
20.	total_ct_chng_q4_q1	Numeric	Change in the number of transactions between Q4 and Q1
21.	avg_utilization_ratio	Numeric	Average utilization ratio of the credit card
22.	naive_bayes_*1	Categorical	To be dropped as suggested by dataset author
23.	naive_bayes_*2	Categorical	To be dropped as suggested by dataset author

4.3 Data Exploration and Feature Trend Analysis

Based on the exploratory analysis performed on the dataset, several observations can be made. Firstly, it is seen that only 16% out of the ~10,000 customers have actually churned, making the classes highly imbalanced. This is essential to note because if 84% customers didn't churn, and our model return every single prediction as "not churned", it will still be 84% accurate which is clearly misleading. To counteract this imbalance, we use two techniques. One, we apply SMOTE for class rebalancing, and second, we leverage the confusion matrix along with the f1-score, precision, and recall to get a better understanding of the model performance instead of relying solely on the accuracy metric.

There are no missing values in the dataset, eliminating the need for any missing value imputations to be performed. The mean age for all customers is around 46, and we observe a bell-curve distribution for this feature. There are almost the same proportion of male and female customers, with the female population being slightly higher. In terms of education, the highest proportion of customers are graduates, with high-school, uneducated, and unknown categories following them. A significant chunk of the population makes less than $40k per annum which also makes the biggest proportion of people in the dataset. A mean credit limit across the dataset at $8,631 with an average of 30% utilization. In terms of customer tenure, a sharp peak around 36 months is observed, with 13 months as the minimum and 56 months as the maximum. This is relevant in ensuring that only the customers who have stayed for at-least a year are used to train the model. The mean of the number of dependents is 2.3 supports the observation that approximately half the population is married.

The features Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1 capture changes in transaction amounts and counts between the fourth and first quarters. Significant positive changes could indicate renewed interest or increased engagement, while declines may signal disengagement and a higher likelihood of churn. These metrics help capture behavioral shifts that may precede customer attrition. To identify relationships between features, a correlation heatmap was generated for numerical features. Strong correlations were observed between Total_Trans_Amt and Total_Trans_Ct, suggesting that customers with higher transaction counts also tend to have higher transaction amounts. This correlation might indicate that certain customer segments have both high engagement and spending patterns, which could be associated with loyalty. Other features showed weaker correlations with attrition, indicating a more complex, multi-factor relationship. For example, Avg_Utilization_Ratio may have a slight correlation with Credit_Limit, suggesting that customers with higher limits use a smaller percentage of their available credit, which might indicate greater financial stability.

4.4 Data Preprocessing

The entire dataset is processed to make it suitable for model training. It is observed that there are no missing values across all features eliminating the need for imputations strategies. There are two columns pertaining to the naïve bayes classification in the dataset that are dropped as recommended by the dataset author. Further, we also drop the column with the client numbers because it is neither a dependent or an independent feature. The categorical features, namely 'Attrition_Flag', 'Gender', 'Education_Level', 'Marital_Status', 'Income_Category', and 'Card_Category' are converted to numerical values for effective model training. To ensure

convergence for the Logistic Regression algorithm, numerical scaling is applied to standardize all numerical feature values, essentially centering them around 0 with a unit variance. This ensures that each feature contributes equally in magnitude, which can improve model performance, especially in algorithms sensitive to feature scales like Logistic Regression. With the preliminary preprocessing done, a 70-30 split is applied on the dataset for training and testing, respectively. Finally, a significant class imbalance is seen with only 16% customers actually churning which is counteracted by oversampling the training dataset with SMOTE (Synthetic Minority Over-sampling Technique) such that there are an equal number of records for both output classes. Going forward, the class representing the churned customers is referred to as the positive class, and the population that stayed as the negative class. After rebalancing of training data, the class counts go from 5,957 for the negative class and 1,131 for the positive class to 5,957 for the both the positive and negative classes. Because SMOTE generates synthetic data, albeit in a controlled manner, no oversampling is performed on the testing split to ensure accurate results in the performance metrics.

CHAPTER 5: EXPERIMENTAL MODEL EVALUATION

This chapter presents the experimental results of the different modeling and XAI techniques applied on the dataset as part of the study, which includes Logistic Regression, Decision Tree, Explainable Boosting Machine, and Random Forest. These are then evaluated in the subsequent chapter where the conclusion is drawn and discussed.

5.1 Logistic Regression

Logistic Regression (LR) is a well-known machine learning technique suitable for binary classification tasks on structured and tabular data. Being a linear model, it lacks complexity and may not generalize well on complex problem statements, it outshines other techniques when it comes to computational cost for training.

Table 2: Confusion Matrix for Logistic Regression

	Reality: Churned (Positive Class)	**Reality: Not-churned** (Negative Class)
Predicted: Churned (Positive Class)	**403** (True Positive)	**348** (False Positive)
Predicted: Not-churned (Negative Class)	**93** (False Negative)	**2195** (True Negative)

Confusion Matrix Analysis:

- True Negatives (TN): 2195 non-churned customers correctly classified.
- False Positives (FP): 348 non-churned customers incorrectly classified as churned.
- False Negatives (FN): 93 churned customers incorrectly classified as non-churned.
- True Positives (TP): 403 churned customers correctly classified.

Table 3: Performance Metrics for Logistic Regression

	Precision	**Recall**	**F1-Score**	**Support**
Class 0: Not-churned (Negative Class)	0.96	0.86	0.91	2543
Class 1: Churned (Positive Class)	0.54	0.81	0.65	496
Accuracy			0.85	3039
Macro Average	0.75	0.84	0.78	3039
Weighted Average	0.89	0.85	0.87	3039

Performance Metrics:

- Precision:
    - For class 0 (non-churned): 0.96, indicating that when the model predicts a customer will not churn, it is correct 96% of the time.
    - For class 1 (churned): 0.54, meaning only 54% of predicted churn cases are actual churn cases. This indicates over-prediction of churn cases, consistent with the

higher false positive rate.

- Recall:
    - For class 0: 0.86, meaning the model correctly identifies 86% of non-churned customers.
    - For class 1: 0.81, meaning it correctly identifies 81% of churned customers.
- F1 Score:
    - For class 0: 0.91, a high value showing that the model is well-calibrated for the non-churn class.
    - For class 1: 0.65, indicating a moderate balance between precision and recall for churned customers.

Overall Metrics:

- Support: 2543 for class 0 (non-churned), 496 for class 1 (churned), and 3039 combined.
- Accuracy: 0.85, meaning the model correctly classifies 85% of all cases.
- Macro Average: Precision and recall averages are 0.75 for precision, 0.84 for recall, and 0.78 for F1-Score. These are lower than the weighted average, reflecting the influence of the minority class (churned) on overall performance.
- Weighted Average: The weighted averages are 0.89 for precision, 0.85 for recall, and 0.87 for F1-Score.

While the true positive and true negative figures look promising, the false positives count of 348 is relatively high especially when compared to the support for the positive class at 496 records.

The following visualization depicts the local interpretation of a prediction by the Logistic Regression model. The particular prediction is for an accurately predicted churned customer with the total transaction count being the most important feature leading to the model decision. The total transaction amount and gender have a negative influence on the positive class prediction.

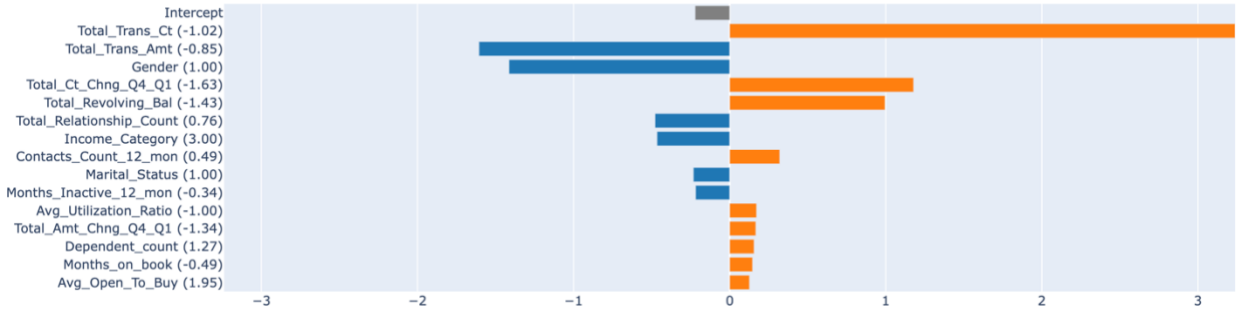

Figure 1: Local Interpretation for Logistic Regression

Similar to the local interpretations, Logistic Regression's inherently explainable nature enables peeking into the general model behavior. The following figure demonstrates the feature importance for the model both in terms of magnitude and class inclination. A longer bar means a higher weightage/coefficient for that feature, which a right-pointed bar represents a higher value for that feature will push the prediction towards the positive class, and a left pointed-bar represents a higher value would decrease the probability of the model predicting the positive class.

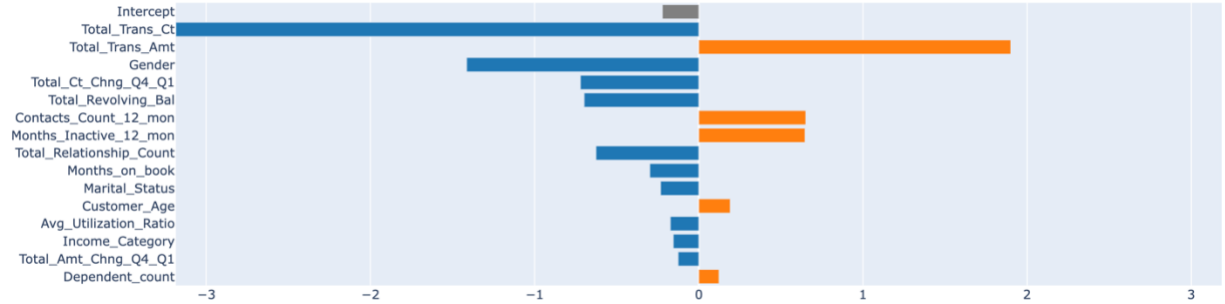

Figure 2: Global Interpretation for Logistic Regression

For the global interpretation, take for instance the total transaction count. It has a large negative coefficient represented by a large leftward bar. This would reasonably and intuitively mean that a higher transaction count will reduce the likelihood of the customer churning as per the model training. Similarly, the large positive coefficient for the Total Transaction Amount suggests that as the amount of money a customer spends increases, the likelihood of churn increases. This could imply that customers with higher transaction amounts may feel they are spending too much or may be more likely to seek alternatives or competitive offers, leading to churn. It is to be noted that these feature importance figures are model specific and represent what the model has learnt basis which it makes the predictions.

Summary: Logistic Regression is simple, fast to train, and interpretable but suffers from limited predictive power on complex data.

5.2 Decision Tree

Decision Trees are inherently explainable models in the form of a binary tree. Each node is a decision where the flow will go either to the left or the right child node. The leaf nodes represent the final prediction. For any model inference, just looking at the path followed in the tree describes exactly why the decision was made. Similarly, a global interpretation of the model is possible by looking at a visualization of the entire decision tree.

Table 4: Confusion Matrix for Decision Tree

	Reality: Churned (Positive Class)	**Reality: Not-churned** (Negative Class)
Predicted: Churned (Positive Class)	**374** (True Positive)	**161** (False Positive)
Predicted: Not-churned (Negative Class)	**122** (False Negative)	**2382** (True Negative)

Confusion Matrix Analysis:

- True Negatives (TN): 2382 non-churned customers correctly classified.
- False Positives (FP): 161 non-churned customers incorrectly classified as churned.
- False Negatives (FN): 122 churned customers incorrectly classified as non-churned.
- True Positives (TP): 374 churned customers correctly classified.

Table 5: Performance Metrics for Decision Tree

	Precision	**Recall**	**F1-Score**	**Support**
Class 0: Not-churned (Negative Class)	0.95	0.94	0.94	2543
Class 1: Churned (Positive Class)	0.70	0.75	0.73	496
Accuracy			0.91	3039
Macro Average	0.83	0.85	0.83	3039
Weighted Average	0.91	0.91	0.91	3039

Performance Metrics:

- Precision:
    - For class 0 (non-churned): 0.95, indicating that when the model predicts a customer will not churn, it is correct 95% of the time.
    - For class 1 (churned): 0.70, meaning 70% of predicted churn cases are actual churn cases. This still indicates some over-prediction of churn cases.
- Recall:
    - For class 0: 0.94, meaning the model correctly identifies 94% of non-churned customers.
    - For class 1: 0.75, meaning it correctly identifies 75% of churned customers.
- F1 Score:
    - For class 0: 0.94, a high value showing that the model is well-calibrated for the

non-churn class.

- For class 1: 0.73, indicating a moderate balance between precision and recall for churned customers.

Overall Metrics:

- Support: 2543 for class 0 (non-churned), 496 for class 1 (churned), and 3039 combined.
- Accuracy: 0.91, meaning the model correctly classifies 91% of all cases.
- Macro Average: Precision and recall averages are 0.83 for precision, 0.85 for recall, and 0.83 for F1-Score. These are lower than the weighted average, reflecting the influence of the minority class (churned) on overall performance.
- Weighted Average: 0.91 for precision, 0.91 for recall, and 0.91 for F1-Score.

The following figure illustrates how a prediction for a churned customer is made by traversing a path in the tree based on the decisions made at each node. The thicker edges between nodes represents that more observations passed through them during training.

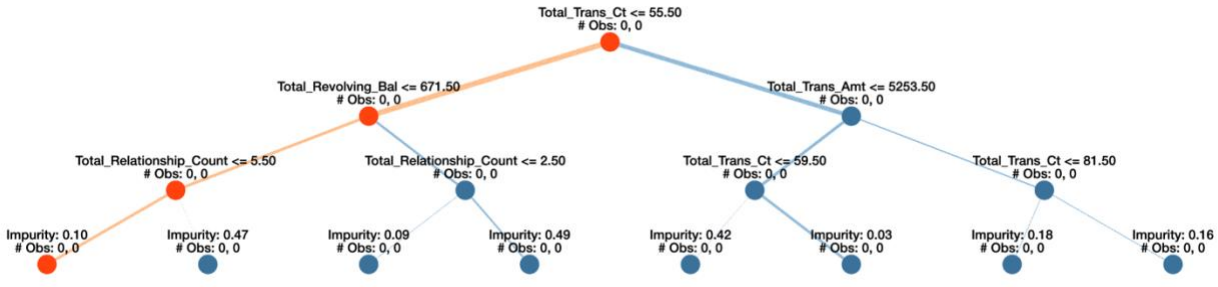

Figure 3: Local Interpretation for Decision Trees

Similarly, it is also possible to view the entire decision tree and not be limited to just local interpretations. The following figure shows a section of the decision tree for which the performance metrics have been discussed above. The complete tree is available in the code repository available in the appendix.

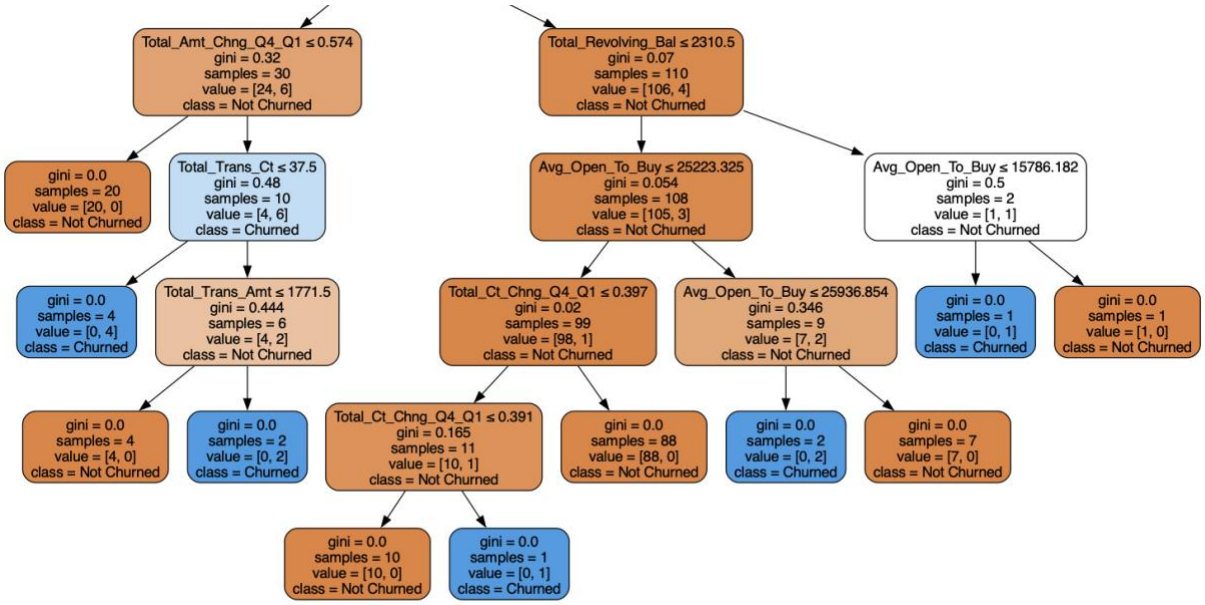

Figure 4: Sectional representation of the complete Decision Tree

Each node is either a decision node or a leaf node. Decision nodes are responsible for checking conditional statements and strictly have two children. If the condition is true, left child is traversed, else the decision flow moves to the right child. Eventually, the final prediction is made at the leaf nodes. In this study, the model was trained such that each leaf node is associated with a class, and the prediction is directly derived from the leaf node that is reached after traversing the decision nodes. For example, in figure 4, the orange nodes represent not-churned customers while the blue

nodes represent churned customers. For the decision nodes, it is obvious that observations for both classes may be flowing through it, but the primary class for the node is still calculated and represented as the majority class. The opacity of the color clearly conveys the majority class, as well as the class mix passing through a node, for example, a light blue node would represent a majority class of churning customers. The class mix is measured by the gini value, also known as the class impurity of a node. This would be 0 for leaf nodes, while the root node is ideally 0.5 because 50% records went to the left and the remaining 50% to the right during model training. The features that are able to split the data effectively have a low entropy and high information gain, and are desirable for effective model training.

Summary: Decision Trees are highly interpretably and intuitive to understand, both for local and global inferences. They are non-linear models so they are able to capture relationships in complex data more effectively than simpler models such as Linear Regression.

5.3 Explainable Boosting Machine (EBM)

EBM is an inherently interpretable model that combines boosting techniques with additive models. EBM can capture non-linear relationships while maintaining transparency, providing local and global explanations for each prediction. It was developed by Microsoft with the primary purpose of achieving the performance of complex models while being inherently explainable like simpler models.

Table 6: Confusion Matrix for Explainable Boosting Machine (EBM)

	Reality: Churned (Positive Class)	**Reality: Not-churned** (Negative Class)
Predicted: Churned (Positive Class)	**448** (True Positive)	**44** (False Positive)
Predicted: Not-churned (Negative Class)	**48** (False Negative)	**2499** (True Negative)

Confusion Matrix Analysis:

- True Negatives (TN): 2499 non-churned customers correctly classified.
- False Positives (FP): 44 non-churned customers incorrectly classified as churned.
- False Negatives (FN): 48 churned customers incorrectly classified as non-churned.
- True Positives (TP): 448 churned customers correctly classified.

Table 7: Performance Metrics for Explainable Boosting Machine (EBM)

	Precision	**Recall**	**F1-Score**	**Support**
Class 0: Not-churned (Negative Class)	0.98	0.98	0.98	2543
Class 1: Churned (Positive Class)	0.91	0.90	0.91	496
Accuracy			0.97	3039
Macro Average	0.95	0.94	0.94	3039
Weighted Average	0.97	0.97	0.97	3039

Performance Metrics:

- Precision:
    - For class 0 (non-churned): 0.98, indicating that when the model predicts a customer will not churn, it is correct 98% of the time.
    - For class 1 (churned): 0.91, meaning 91% of predicted churn cases actual churned.
- Recall:
    - For class 0: 0.98, meaning the model correctly identifies 98% of non-churned customers.
    - For class 1: 0.90, meaning it correctly identifies 90% of churned customers.
- F1 Score:
    - For class 0: 0.98, a high value showing that the model is well-calibrated for the non-churn class.

- For class 1: 0.91, indicating a great balance between precision and recall for churned customers.

Overall Metrics:

- Support: 2543 for class 0 (non-churned), 496 for class 1 (churned), and 3039 combined.
- Accuracy: 0.97, meaning the model correctly classifies 97% of all cases.
- Macro Average: Precision and recall averages are 0.95 for precision, 0.94 for recall, and 0.94 for F1-Score.
- Weighted Average: 0.97 for precision, 0.97 for recall, and 0.97 for F1-Score.

The following figures 5 and 6 represent the local and global interpretations for the trained EBM model.

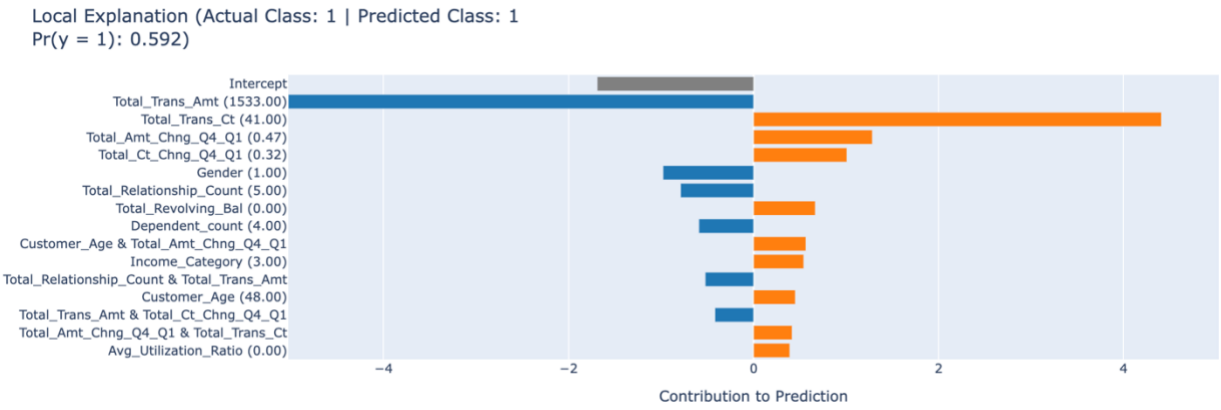

Figure 5: Local interpretation for Explainable Boosting Machine (EBM)

The figure clearly demonstrates using bar lengths, direction, and order to convey why the prediction was made for a customer who was accurately identified to churn.

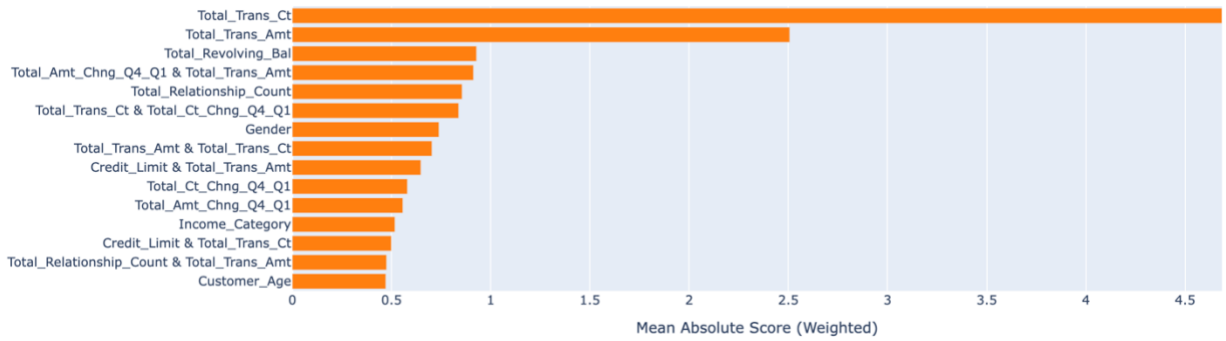

Figure 6: Global feature importance for Explainable Boosting Machine (EBM)

The global feature importance gives a clear picture on what features the model values the most, with the total transaction count leading by a great margin, followed by the total transaction amount and the total revolving balance as the three most important features for the model.

5.4 Random Forest

Random Forest is an ensemble model composed of multiple decision trees. It aggregates the predictions of individual trees to improve accuracy and robustness, but this it is not designed for interpretability but for performance and adapting to more complex data patterns. Therefore, it is often considered a black-box model and we apply LIME and SHAP techniques to the trained Random Forest model to make it interpretable.

Table 8: Confusion Matrix for Random Forest Model

	Reality: Churned (Positive Class)	**Reality: Not-churned** (Negative Class)
Predicted: Churned (Positive Class)	**436** (True Positive)	**66** (False Positive)
Predicted: Not-churned (Negative Class)	**60** (False Negative)	**2477** (True Negative)

Confusion Matrix Analysis:

- True Negatives (TN): 2477 non-churned customers correctly classified.
- False Positives (FP): 66 non-churned customers incorrectly classified as churned.
- False Negatives (FN): 60 churned customers incorrectly classified as non-churned.
- True Positives (TP): 436 churned customers correctly classified.

Table 9: Performance Metrics for Random Forest Model

	Precision	**Recall**	**F1-Score**	**Support**
Class 0: Not-churned (Negative Class)	0.98	0.97	0.98	2543
Class 1: Churned (Positive Class)	0.87	0.88	0.87	496
Accuracy			0.96	3039
Macro Average	0.92	0.93	0.92	3039
Weighted Average	0.96	0.96	0.96	3039

Performance Metrics:

- Precision:
    - For class 0 (non-churned): 0.98, indicating that when the model predicts a customer will not churn, it is correct 98% of the time.
    - For class 1 (churned): 0.87, meaning 87% of predicted churn cases actual churned.
- Recall:
    - For class 0: 0.97, meaning the model correctly identifies 97% of non-churned customers.
    - For class 1: 0.88, meaning it correctly identifies 88% of churned customers.
- F1 Score:
    - For class 0: 0.98, a high value showing that the model is well-calibrated for the non-churn class.

- For class 1: 0.87, indicating a good balance between precision and recall for churned customers.

Overall Metrics:

- Support: 2543 for class 0 (non-churned), 496 for class 1 (churned), and 3039 combined.
- Accuracy: 0.96, meaning the model correctly classifies 96% of all cases.
- Macro Average: Precision and recall averages are 0.92 for precision, 0.93 for recall, and 0.92 for F1-Score.
- Weighted Average: 0.96 for precision, 0.96 for recall, and 0.96 for F1-Score.

Although the model achieves great performance, it is not interpretable. In the following figure 7, LIME is applied to get the local interpretation of a sample prediction.

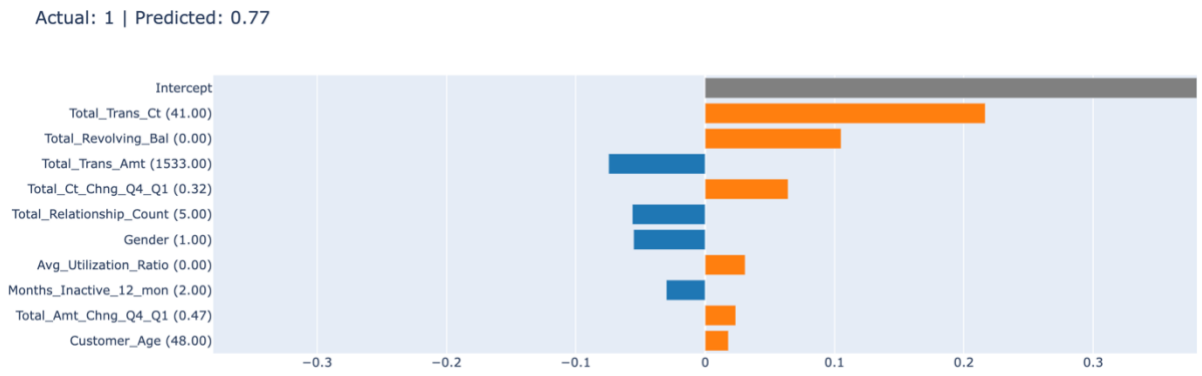

Figure 7: LIME for local interpretability of a black-box model (Random Forest)

By leveraging LIME technique, an estimation of the feature importance was made on the chosen black-box model (Random Forest). LIME uses a surrogate model which is a simple linear model

such as logistic regression to make this estimation. Small perturbations are introduced to the input features to analyze its effect on the output prediction of the complex black-box model, essentially giving us a window into the importance of the different features for a particular prediction. However, this technique of gaining insights has a limitation of only allowing local interpretations which is where SHAP can help to understand the general model behavior not centered around a particular prediction as shown in the following figure 8.

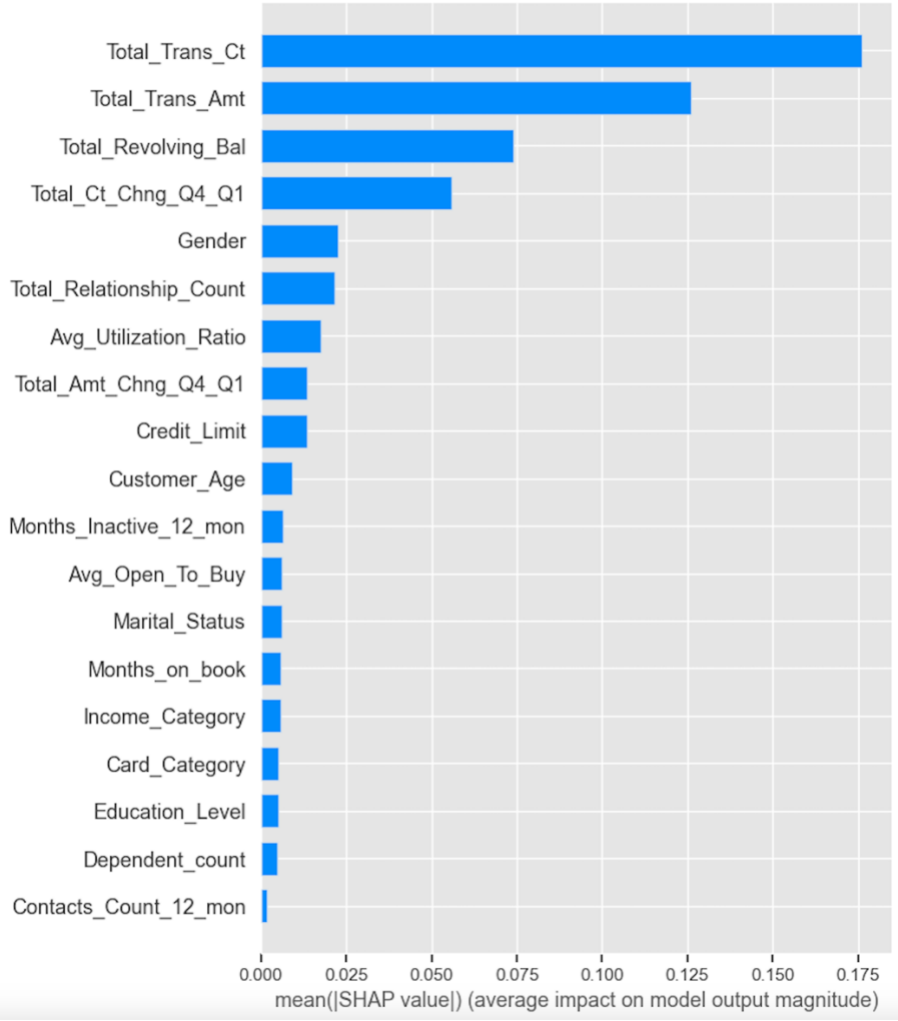

Figure 8: SHAP for global interpretability of the Random Forest model

CHAPTER 6: CONCLUSION AND DISCUSSION

In this study, we explored and evaluated multiple machine learning models to predict customer churn, focusing on balancing accuracy and interpretability. Each model demonstrated unique strengths and limitations, leading to insights on when and why certain models might be more suitable based on the business requirements.

Logistic Regression provided a quick and simple baseline, with training taking only seconds. However, its generalization ability was limited, as reflected in a significant number of false positives and false negatives. Although it achieved an accuracy of 85%, its recall and precision were not optimal for the positive class, indicating it might not be ideal for complex churn prediction tasks. Nevertheless, its inherent interpretability made it easy to understand how features influenced predictions, which could be beneficial in simpler or highly linear cases.

The Decision Tree model showed improved accuracy (91%) compared to Logistic Regression and provided a clear structure for interpretability through tree visualization. This made it easy to understand the decision-making process for churn prediction. The tree's branches allowed us to trace how each feature affected outcomes, instilling confidence in its interpretability. However, while it performed well overall, it still struggled with precision and recall for the positive (churned) class, suggesting limitations in handling imbalanced datasets.

Explainable Boosting Machine (EBM) achieved the highest accuracy (97%) and offered both local and global explanations, making it highly interpretable. It provided valuable insights into feature

importance, allowing a nuanced understanding of how individual factors influenced churn predictions. Despite its impressive performance, EBM is computationally intensive and took significantly longer to train. Additionally, its suitability is limited to structured data, meaning it may not be ideal for image or natural language processing (NLP) tasks without extensive preprocessing to convert unstructured data into a structured format.

Random Forest performed similarly to EBM in accuracy, with a score of 96%. However, due to its complexity, we needed model-agnostic explainability techniques like LIME and SHAP to interpret its predictions. Although these techniques provided meaningful insights into feature importance, they required additional computational resources and time, particularly SHAP, which took longer to compute. Despite these efforts, Random Forest's interpretability remains less intuitive than that of EBM or simpler models like Decision Trees.

Model-agnostic tools like LIME and SHAP were invaluable in enhancing interpretability for the black-box model, and they can be applied to a range of model types, including Random Forest, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks. They generate visual explanations suited to different data types, such as feature importance for tabular data, heatmaps for images, and token importance for text. However, their use is generally reserved for situations where interpretability is crucial, as these tools add computational overhead. Additionally, these techniques may struggle with more complex, generative models like generative AI such as chatbots and image generators, where outcomes involve high-dimensional interactions and complex dependencies.

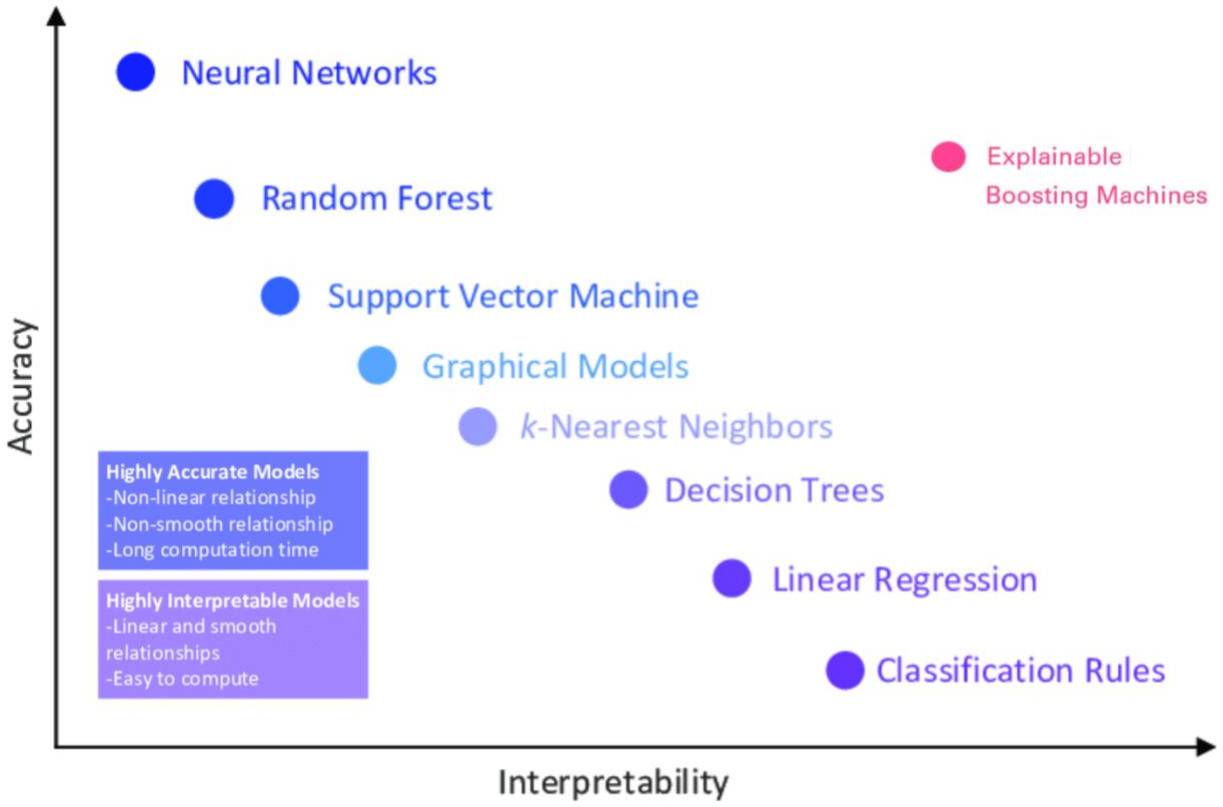

Figure 9: Plot of Model Accuracy vs. Interpretability

In conclusion, our findings suggest that for straightforward, linear tasks, simpler models like Logistic Regression or Decision Trees are advantageous due to their interpretability and low computational cost. For more complex but structured data, EBM offers a powerful solution, combining high accuracy with inherent interpretability. However, in cases involving unstructured data or domains where EBM is unsuitable (e.g., image classification), more specialized models such as neural networks may be preferred, accompanied by LIME and SHAP to enhance transparency. Ultimately, the choice of model depends on the specific trade-off between interpretability, accuracy, and computational cost, as visualized in our accuracy-interpretability spectrum diagram. This study underscores the importance of selecting models not only for predictive power but also for transparency.

REFERENCES

1. Adadi, Amina, and Mohamed Berrada. “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI).” IEEE Access 6 (2018).
2. Gunning, David. “Explainable artificial intelligence (XAI).” Defense Advanced Research Projects Agency (DARPA), 2017.
3. Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint arXiv:1702.08608 (2017).
4. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why should I trust you? Explaining the predictions of any classifier.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
5. Lipton, Zachary C. “The mythos of model interpretability.” Communications of the ACM 61, no. 10 (2018).
6. Molnar, Christoph. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019.
7. Hasan, Md Rokibul & Gazi, Md & Gurung, Nisha. “Explainable AI in Credit Card Fraud Detection: Interpretable Models and Transparent Decision-making for Enhanced Trust and Compliance in the USA.” Journal of Computer Science and Technology Studies 6 (2024).
8. Saranya, A., and R. Subhashini. “A systematic review of Explainable Artificial Intelligence models and applications.” Decision Analytics Journal 9 (2023).
9. Angelov, P. P., et al., “Explainable artificial intelligence: an analytical review.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11(5) (2021).

10. Hassija, V., Chamola, V., Mahapatra, A. *et al.* “Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence.” *Cogn Comput* 16, 45–74 (2024).

APPENDIX

1. GitHub Code Repository: https://harshitjindal.com/xai-dissertation
2. Dataset available at: Kaggle Link

Checklist

This checklist is to be attached as the last page of the final report.

This checklist is to be duly completed, verified and signed by the student.

1.	Is the final report neatly formatted with all the elements required for a technical Report?	Yes
2.	Is the Cover page in proper format as given in Annexure A?	Yes
3.	Is the Title page (Inner cover page) in proper format?	Yes
4.	(a) Is the Certificate from the Supervisor in proper format? (b) Has it been signed by the Supervisor?	Yes Yes
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	Yes Yes
6.	Is the title of your report appropriate? The title should be adequately descriptive, precise and must reflect scope of the actual work done. Uncommon abbreviations / Acronyms should not be used in the title	Yes
7.	Have you included the List of abbreviations / Acronyms?	Yes
8.	Does the Report contain a summary of the literature survey?	Yes
9.	Does the Table of Contents include page numbers? i. Are the Pages numbered properly? (Ch. 1 should start on Page # 1) ii. Are the Figures numbered properly? (Figure Numbers and Figure Titles	Yes Yes

	should be at the bottom of the figures) iii. Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) iv. Are the Captions for the Figures and Tables proper? v. Are the Appendices numbered properly? Are their titles appropriate	Yes Yes Yes Yes
10.	Is the conclusion of the Report based on discussion of the work?	Yes
11.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report?	Yes Yes Yes
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a PowerPoint Presentation, or a user manual. Source code of software need not be included in the report.	Yes

Declaration by the student:

I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

_____

Signature of the Student

Harshit Jindal

2022MT93524

November 18, 2024