Università di Catania

Department of Economics and Business

Master's Degree in Data Science

Explainable AI: The SHAP Algorithm

MASTER DEGREE THESIS

Author
Asad Aslam

Supervisor
Prof. Gallo giovanni

ACADEMIC YEAR 2023/2024
September 25th, 2024

Asad Aslam

Explainable AI: The SHAP Algorithm

Master's Degree Thesis

UNIVERSITY OF CATANIA

September 2024

Contents

List of Figures

List of Tables

List of Abbreviations

XAI	**E**xplainable**A**ritifical **I**ntelligence
SHAP	**SH**apley **A**dditive **e**xPlanations
LIME	**L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations

Chapter 1

Introduction

1.1 Background Motivation

In recent years, the proliferation of machine learning models across diverse industries has significantly enhanced decision-making processes and predictive capabilities. However, as these models become increasingly complex, their inherent opacity presents challenges in understanding how they arrive at their predictions. This lack of transparency can hinder user trust, limit adoption, and raise ethical concerns, particularly in critical domains such as healthcare, finance, and criminal justice. Explainable Artificial Intelligence (XAI) has emerged as a pivotal area of research aimed at addressing the interpretability and transparency of machine learning models. By providing human-understandable explanations for model predictions, XAI techniques enable stakeholders to comprehend and trust the decisions made by AI systems. Among various XAI methods, SHAP (SHapley Additive exPlanations) has garnered significant attention for its effectiveness in elucidating the inner workings of black-box models. The primary motivation behind this thesis is to delve into the intricacies of the SHAP technique, understand its underlying principles, and evaluate its applicability in real-world scenarios. By conducting a comprehensive study, we aim to deepen our understanding of XAI methodologies, explore the strengths and limitations of SHAP, and assess its performance across diverse datasets and machine learning models.

1.2 Research Objectives

The main objectives of this thesis are as follows:

1. Deeply understand the theoretical foundations and operational mechanisms of SHAP.

2. Investigate the effectiveness of SHAP in providing interpretable explanations for machine learning models across various domains.

3. Evaluate the robustness and scalability of SHAP on real-world datasets with different characteristics, such as size, dimensionality, and complexity.

4. My colleague is conducting a similar study using LIME (Local Interpretable Model-agnostic Explanations), another widely-used XAI technique. Together, we will compare the performance of SHAP and LIME in terms of interpretability, computational efficiency, and ease of implementation.

5. Demonstrate the practical utility of SHAP through a hands-on project implemented in Python, showcasing their application in real-world scenarios.

Chapter 2

Literature Overview

2.1 Explainable Artificial Intelligence (XAI)

XAI stands for Explainable Artificial Intelligence. It refers to the concept of designing and developing AI systems in such a way that their decisions and outputs can be easily understood and explained by humans. The need for XAI arises because many modern AI algorithms, particularly those based on deep learning and neural networks, often operate as black boxes, meaning their decision-making processes are not transparent or interpretable by humans. Here are some key points about XAI:

1. Transparency: XAI aims to increase the transparency of AI systems, allowing users to understand how a decision was made, what factors contributed to it, and why a certain output was produced.

2. Interpretability: XAI techniques seek to make AI models more interpretable by humans. This involves techniques such as feature importance analysis, attention mechanisms, and generating human-readable explanations for model predictions.

3. Trust: By making AI systems more understandable, XAI can help build trust between users and AI systems. Users are more likely to trust and adopt AI systems if they can understand how they work and why they make certain decisions.

4. Regulatory Compliance: In some domains, such as healthcare and finance, there are regulatory requirements for transparency and accountability in decision-making processes. XAI techniques can help AI systems comply with these regulations.

5. Bias and Fairness: XAI can help identify and mitigate bias in AI systems by providing insights into how decisions are made and which factors influence them. This can help ensure that AI systems are fair and unbiased across different demographic groups.

6. Education and Collaboration: XAI promotes collaboration between AI developers, domain experts, and end-users by facilitating communication and understanding of AI systems. It also helps educate users about the capabilities and limitations of AI technology.

Overall, XAI plays a crucial role in making AI systems more accountable, trustworthy, and understandable, thus enabling their broader adoption and acceptance in various domains. "Black box" and "glass box" are terms used to describe different levels of transparency and understanding in AI systems.

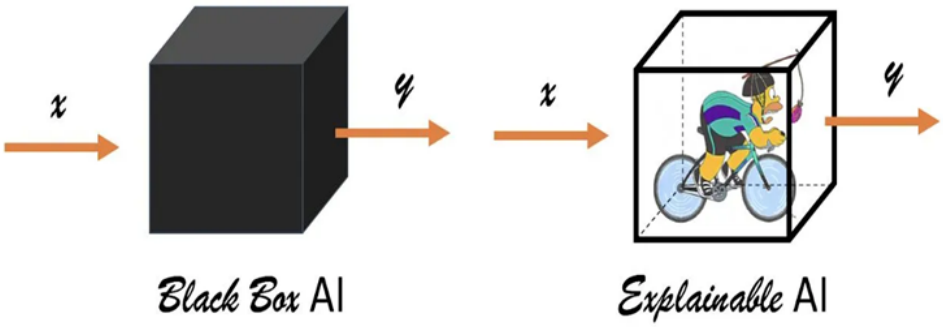

FIGURE 2.1: Black box AI vs Explainable AI

2.2 Black Box

In the context of AI, a black box refers to a system whose internal workings are opaque or not easily understandable to humans. This means that although the inputs and outputs of the system may be known, the process by which the outputs are generated from the inputs is not transparent. Many deep learning models, especially complex neural networks with multiple layers, are often considered black boxes because it can be challenging to interpret how they arrive at their decisions or predictions. Black box models may be highly accurate in their predictions but lack interpretability, making it difficult for humans to trust or understand their decisions.

2.2.1 Black Box Algorithms

1. Deep Learning Neural Networks.
2. Ensemble Methods (e.g., Random Forest, Gradient Boosting).
3. Support Vector Machines (SVM).

2.3 Glass Box

In contrast, a glass box (sometimes also referred to as a white box or transparent box) is a system whose internal workings are transparent and easily understandable to humans. Glass box models are designed in such a way that the decision-making process is clear and interpretable. This often involves using simpler algorithms or models that prioritize transparency over complexity. While glass box models may sacrifice some predictive accuracy compared to

black box models, they offer the advantage of increased interpretability, which can be valuable in contexts where understanding the reasoning behind decisions is important.

2.3.1 Glass Box Algorithms

1. Linear Regression.
2. Decision Trees.
3. Logistic Regression.
4. Naive Bayes.

While this categorization provides a general guideline, it's essential to note that interpretability can vary within algorithm types depending on factors such as hyperparameters, model complexity, and feature engineering. Additionally, techniques such as feature importance analysis and model-agnostic interpretability methods can sometimes enhance the interpretability of black box models to some extent. I am going to use Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machines (SVM). Let's categorize in terms of their interpretability.

Decision trees are generally considered glass box models. They are highly interpretable because they make decisions based on a series of simple rules. Each node in the tree represents a decision based on a feature, and the paths from the root to the leaves represent decision-making processes. This transparency makes decision trees easy to understand and interpret.

Random Forest, which is an ensemble method, is typically considered a black box model. While each individual decision tree in the forest is interpretable, the combination of many trees makes the overall model more complex and less interpretable. Random Forest models are known for their high predictive accuracy but lack the transparency of single decision trees.

Gradient Boosting, similar to Random Forest, is an ensemble method that combines multiple weak learners to form a strong learner. While Gradient Boosting models can provide high accuracy, especially in structured data tasks, they are generally considered black box models. The combination of decision trees through boosting techniques increases model complexity, making it harder to interpret compared to single decision trees.

SVMs are typically considered black box models. While they can produce accurate predictions, especially in high-dimensional spaces, the decision boundaries learned by SVMs can be complex and difficult to interpret, particularly in cases where non-linear kernels are used. SVMs focus on finding the optimal hyperplane that separates classes in the feature space, and understanding the decision-making process may require additional effort compared to more transparent models like decision trees.

Transforming black box models into Explainable Artificial Intelligence (XAI) involves various techniques and approaches aimed at increasing transparency and interpretability. While there's no one-size-fits-all solution, here are some methods commonly used to make black box models more explainable.

Identify and rank the importance of input features that contribute most to the model's predictions. Techniques such as permutation feature importance, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) can provide insights into which features influence the model's decisions.

Provide explanations for individual predictions rather than the entire model. Methods like LIME and SHAP can generate local explanations by approximating the model's behavior around specific instances, making it easier to understand why a particular prediction was made.

2.4 SHAP

SHAP, an acronym for SHapley Additive exPlanations, is a widely-used method in the field of machine learning interpretability that provides a consistent and reliable framework for explaining individual predictions. SHAP's power lies in its ability to assign each feature in a model an importance value for a particular prediction, offering detailed insights into the inner workings of complex models. This method is rooted in the principles of cooperative game theory, specifically drawing from the concept of Shapley values, a solution introduced by economist Lloyd Shapley to fairly distribute the total gains (or payoffs) among players in a cooperative game based on their contributions. In the context of machine learning, SHAP considers features as "players" and predictions as "payoffs." The method systematically evaluates how each feature contributes to a specific prediction by determining the impact of including or excluding it. This approach allows SHAP to provide a unified and consistent explanation across different models, regardless of their complexity or underlying structure. Whether dealing with simple linear models or intricate deep neural networks, SHAP offers a robust solution for model interpretability, ensuring that users can understand and trust the predictions made by these models. SHAP's growing popularity can be attributed to its flexibility and compatibility with a wide range of machine learning models, including those considered as black-box models due to their opaque decision-making processes. By translating these complex models into interpretable outputs, SHAP has become a crucial tool in various domains, from healthcare to finance, where understanding the reasoning behind a model's predictions is not just beneficial but often essential.

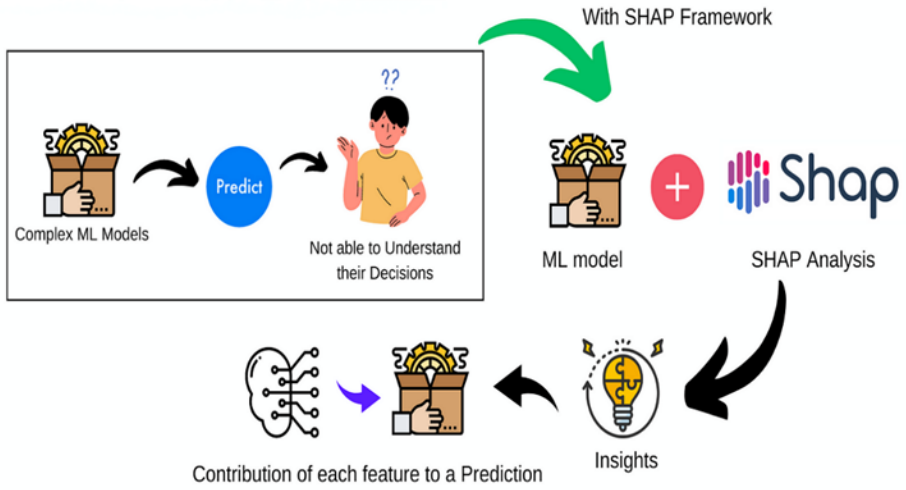

FIGURE 2.2: simplified overview of how SHAP works

2.4.1 Calculate Shapley Values

The calculation of Shapley values is central to SHAP's functionality. Shapley values represent the contribution of each feature to the difference between the model's output and the average output across all predictions. This calculation involves considering all possible combinations of features, assessing the model's output with and without each feature, and then averaging these contributions across all possible subsets of features. For each feature, SHAP calculates its Shapley value by evaluating its marginal contribution across all possible coalitions of features. This is a computationally intensive process, particularly for models with a large number of features, as it requires the evaluation of the model's prediction for every possible subset of features. However, this exhaustive approach ensures that the resulting Shapley values accurately reflect the true importance of each feature in the context of the specific prediction being analyzed. By computing Shapley values in this manner, SHAP ensures that each feature's contribution is fairly assessed, taking into account the complex interactions and dependencies that may exist between features. This is particularly important in models where features are not independent and may have synergistic effects when considered together.

2.4.2 Feature Importance

One of the key outcomes of SHAP analysis is the generation of feature importance scores based on Shapley values. These scores provide a quantitative measure of the extent to which each feature influences a particular prediction. Positive Shapley values indicate that a feature contributes positively to the prediction, increasing the likelihood of the predicted outcome, while negative values suggest that a feature contributes negatively, reducing the likelihood of the outcome. Feature importance analysis using SHAP is especially valuable in scenarios where model transparency is critical. For example, in healthcare, understanding which factors are driving a diagnosis can help clinicians make more informed decisions, ensuring that the model's predictions align with clinical reasoning. Similarly, in finance, feature importance scores can

help explain why a loan application was approved or denied, providing transparency and helping to build trust in automated decision-making systems. Moreover, by comparing feature importance scores across different instances, SHAP allows for the identification of consistent patterns and trends in the data. This can reveal underlying relationships between features and the target variable, providing valuable insights that can inform feature selection, model improvement, and data-driven decision-making.

2.4.3 Explainability

Explainability is at the heart of SHAP's value proposition. By breaking down complex model predictions into understandable components, SHAP provides users with clear explanations of why a particular prediction was made. These explanations are grounded in the Shapley values, which quantify the contribution of each feature to the model's output. SHAP explanations are particularly useful for addressing the "black-box" nature of many modern machine learning models. In traditional black-box models, such as deep neural networks or ensemble methods, the decision-making process is often opaque, making it difficult for users to understand how the model arrived at a particular prediction. SHAP overcomes this challenge by translating the model's complex internal processes into a set of human-interpretable feature contributions, enabling users to gain insights into the model's behavior. These explanations can be crucial in sensitive applications where model decisions have significant consequences. For instance, in the legal domain, where algorithms may be used to assist in judicial decision-making, it is essential to understand the rationale behind each prediction to ensure that the model's outputs are fair and just. Similarly, in healthcare, where models may be used to recommend treatments or diagnose conditions, understanding the factors that influenced the model's decision can help ensure that the recommendations are appropriate and safe.

2.4.4 Global Interpretability

While SHAP excels at providing explanations for individual predictions, it also offers powerful tools for understanding the overall behavior of a model through global interpretability. By aggregating Shapley values across multiple instances, SHAP can provide insights into the model's general trends, biases, and feature interactions. Global interpretability is achieved by analyzing the distribution of Shapley values across the entire dataset or a representative subset of it. This analysis can reveal which features are most influential on average, highlight potential biases in the model, and uncover complex interactions between features. For example, in a credit scoring model, global interpretability might reveal that certain demographic features consistently have a significant impact on predictions, prompting further investigation to ensure that the model is not inadvertently biased. Global interpretability is essential for gaining a holistic understanding of a model's behavior, particularly in applications where the model is used to make decisions at scale. By providing insights into how the model operates across different scenarios, SHAP helps users identify areas where the model may need improvement, ensuring that it performs reliably and fairly in practice. SHAP's ability to offer both local (instance-level) and global (model-level) interpretability makes it

a versatile tool for model explainability. This dual capability allows users to drill down into specific predictions when needed while also gaining a broader understanding of the model's overall behavior, making SHAP an invaluable resource for both model development and deployment.

2.4.5 SHAP in Practice

SHAP has found widespread application across various industries and domains due to its robustness and versatility. Its ability to provide interpretable insights into complex models has made it a popular choice for a wide range of tasks, from healthcare diagnostics to financial risk assessment and beyond. In healthcare, SHAP has been used to explain predictions made by models for disease diagnosis, treatment recommendations, and patient outcome predictions. By providing transparent explanations, SHAP helps healthcare professionals trust and validate the model's recommendations, ultimately leading to better patient care. In finance, SHAP is commonly used to explain models used for credit scoring, fraud detection, and investment analysis. Financial institutions rely on SHAP to ensure that their models are fair, transparent, and compliant with regulatory requirements, particularly in areas where decision-making must be explainable to customers and regulators. In natural language processing (NLP), SHAP has been applied to explain models used for tasks such as sentiment analysis, text classification, and language translation. By revealing which words or phrases are most influential in a model's predictions, SHAP helps researchers and practitioners understand the linguistic patterns that the model has learned, leading to better model development and refinement. In image classification, SHAP has been used to explain deep learning models that classify images into different categories. By identifying which parts of an image are most influential in the model's prediction, SHAP helps users understand how the model is interpreting visual data, which is particularly useful for tasks such as medical image analysis and autonomous driving.

2.5 SHAP Values

Shapley values, named after the economist Lloyd Shapley who introduced them in cooperative game theory, are a concept used to fairly distribute the value generated by cooperation among a group of players. They provide a way to fairly distribute the payoff among the players based on their individual contributions to the cooperative outcome. In the context of machine learning and model explainability, Shapley values are used to attribute the contribution of each feature to a model's prediction. They quantify the marginal contribution of a feature to the prediction by considering all possible combinations of features and their respective predictions. The key idea is to measure the impact of including a feature in a model's prediction compared to excluding it. Here's a simplified explanation of how Shapley values work:

2.5.1 Consider All Feature Combinations

The calculation of Shapley values involves considering every possible combination of features, which is a key aspect of their fairness. For each feature, the

Shapley value is computed by evaluating its contribution to the model's prediction in every possible coalition of features. This exhaustive consideration ensures that the Shapley value accurately reflects the feature's true importance, taking into account the complex interactions that may exist between features. For example, in a model with three features (A, B, and C), the Shapley value for feature A would be calculated by evaluating the model's prediction with and without feature A in every possible subset of features: B, C, A, B, A, C, and A, B, C. By considering all these subsets, the Shapley value accounts for the different ways in which feature A can interact with the other features, providing a fair assessment of its contribution. The requirement to consider all feature combinations makes the calculation of Shapley values computationally expensive, especially for models with a large number of features. However, this thorough approach is essential for ensuring that the resulting Shapley values are fair and accurate.

2.5.2 Weighted Average of Marginal Contributions

Once all possible feature combinations have been considered, the Shapley value for each feature is calculated as the weighted average of its marginal contributions across all these combinations. The weighting is determined by the number of coalitions that include the feature, ensuring that the Shapley value fairly reflects the feature's importance across different scenarios.

2.5.3 Fair Distribution of Credit

The primary motivation behind using Shapley values in machine learning is their ability to fairly distribute credit for a model's prediction among the input features. By considering all possible combinations of features and averaging the marginal contributions, Shapley values ensure that each feature's contribution is assessed in a balanced and unbiased manner. This fair distribution of credit is particularly important in scenarios where the stakes are high, such as in legal or financial decision-making. In these contexts, it is essential to ensure that the model's predictions are not unduly influenced by any single feature or set of features, and that the contributions of all relevant factors are accurately represented. Moreover, the fairness of Shapley values makes them a valuable tool for auditing and validating machine learning models. By providing a transparent and principled way to attribute importance to features, Shapley values help ensure that models are making decisions based on sound reasoning, rather than being influenced by spurious correlations or biases in the data.

The formula to calculate the shapely Value Øi for feature i is as follows:
Øi = SN{i |S|!(|N|-|S|-1)!/|N|! [f (S U i) f(S)]
where

- N is the set of all features.
- S is a subset of N excluding feature i.
- f(S) is the model's prediction when considering only the features in set S.
- |S| denotes the number of features in set S.
- |N| denotes the total number of features.

2.6 Advantages of SHAP

1. Model Agnostic: One of the significant advantages of SHAP is its model-agnostic nature. It can be applied to any machine learning model, whether it's a linear regression, decision tree, random forest, gradient boosting, or deep neural network. This flexibility makes SHAP widely applicable across various domains and model architectures.

2. Individualized Explanations: SHAP provides explanations at the individual prediction level, offering insights into why a particular prediction was made. This granularity allows users to understand the model's decision-making process on a case-by-case basis, which is crucial for tasks requiring interpretability, such as healthcare or finance.

3. Feature Importance Analysis: By calculating Shapley values for each feature, SHAP quantifies the importance of features in model predictions. This allows users to identify which features are most influential in driving the model's decisions, aiding feature selection, and understanding the underlying relationships in the data.

4. Global Interpretability: In addition to explaining individual predictions, SHAP can aggregate Shapley values across multiple instances to provide insights into the overall behavior of the model. This global interpretability helps users understand the model's general trends, biases, and feature interactions.

5. Handles Complex Models: SHAP can handle complex, black-box models effectively. Even models with high non-linearity or interactions between features can be explained using SHAP. This is particularly valuable in scenarios where model interpretability is crucial, but the model's complexity poses challenges for traditional interpretation methods.

6. Fair Attribution: SHAP offers a principled approach to attributing the model's output to individual features, ensuring fair credit distribution. By considering all possible feature combinations and their contributions, SHAP avoids biases in feature importance attribution, providing a more reliable explanation of model predictions.

7. Wide Application: SHAP has been successfully applied across various domains, including healthcare, finance, natural language processing, image classification, and more. Its versatility and effectiveness make it a popular choice for model explainability and interpretability in diverse applications.

Chapter 3

Experiment

3.1 Student Dropout Dataset

The dataset aims to contribute to reducing academic dropout and failure in higher education by leveraging machine learning techniques. By identifying students at risk of dropping out early in their academic journey, the goal is to implement supportive strategies to enhance their academic success. The dataset includes various features related to the students' academic paths, demographics, and socio-economic factors, known at the time of enrollment. Below is a detailed description of the dataset.

3.1.1 General Information

- Source: The dataset is sourced from the UCI Machine Learning Repository.
- Type: Classification
- Features: 36
- Target Variable: The problem is formulated as a three-category classification task (dropout, enrolled, and graduate) at the end of the normal course duration.
- Missing Values: No missing values are reported.

3.1.2 Features Overview

Feature Name	Type	Demographic	Description	Values/Units	Missing Values
Marital Status	Integer	Marital Status	Marital status of the student	1 – single, 2 – married, 3 – widower, 4 – divorced, 5 – facto union, 6 – legally separated	No
Application mode	Integer		Mode of application	Multiple values (e.g., 1 - 1st phase - general contingent, 2 - Ordinance No. 612/93, etc.)	No
Application order	Integer		Application order (0 - first choice; 9 - last choice)	0-9	No
Course	Integer		The course the student is enrolled in	Multiple values (e.g., 33 - Biofuel Production Technologies, etc.)	No
Daytime/evening attendance	Integer		Attendance time	1 – daytime, 0 - evening	No
Previous qualification	Integer	Education Level	Education level prior to enrollment	Multiple values (e.g., 1 - Secondary education, 2 - Higher education - bachelor's degree, etc.)	No
Previous qualification (grade)	Continuous		Grade of previous qualification	0-200	No
Nacionality	Integer	Nationality	Nationality of the student	Multiple values (e.g., 1 - Portuguese, 2 - German, etc.)	No
Mother's qualification	Integer	Education Level	Education level of the student's mother	Multiple values (e.g., 1 - Secondary Education, etc.)	No
Father's qualification	Integer	Education Level	Education level of the student's father	Multiple values (e.g., 1 - Secondary Education, etc.)	No
Mother's occupation	Integer	Occupation	Occupation of the student's mother	Multiple values (e.g., 0 - Student, 1 - Representatives, etc.)	No
Father's occupation	Integer	Occupation	Occupation of the student's father	Multiple values (e.g., 0 - Student, 1 - Representatives, etc.)	No
Admission grade	Continuous		Admission grade	0-200	No
Displaced	Integer		Whether the student is displaced	1 – yes, 0 – no	No
Educational special needs	Integer		Whether the student has special educational needs	1 – yes, 0 – no	No
Debtor	Integer		Whether the student is a debtor	1 – yes, 0 – no	No
Tuition fees up to date	Integer		Whether the tuition fees are up to date	1 – yes, 0 – no	No
Gender	Integer	Gender	Gender of the student	1 – male, 0 – female	No

Scholarship holder	Integer		Whether the student is a scholarship holder	1 – yes, 0 – no	No
Age at enrollment	Integer	Age	Age of the student at enrollment	Integer	No
International	Integer		Whether the student is an international student	1 – yes, 0 – no	No
Curricular units 1st sem (credited)	Integer		Number of curricular units credited in the 1st semester	Integer	No
Curricular units 1st sem (enrolled)	Integer		Number of curricular units enrolled in the 1st semester	Integer	No
Curricular units 1st sem (evaluations)	Integer		Number of evaluations to curricular units in the 1st semester	Integer	No
Curricular units 1st sem (approved)	Integer		Number of curricular units approved in the 1st semester	Integer	No
Curricular units 1st sem (grade)	Integer		Grade average in the 1st semester	0-20	No
Curricular units 1st sem (without evaluations)	Integer		Number of curricular units without evaluations in the 1st semester	Integer	No
Curricular units 2nd sem (credited)	Integer		Number of curricular units credited in the 2nd semester	Integer	No
Curricular units 2nd sem (enrolled)	Integer		Number of curricular units enrolled in the 2nd semester	Integer	No
Curricular units 2nd sem (evaluations)	Integer		Number of evaluations to curricular units in the 2nd semester	Integer	No
Curricular units 2nd sem (approved)	Integer		Number of curricular units approved in the 2nd semester	Integer	No
Curricular units 2nd sem (grade)	Integer		Grade average in the 2nd semester	0-20	No
Curricular units 2nd sem (without evaluations)	Integer		Number of curricular units without evaluations in the 2nd semester	Integer	No
Unemployment rate	Continuous		Unemployment rate	Percentage	No
Inflation rate	Continuous		Inflation rate	Percentage	No
GDP	Continuous		Gross Domestic Product	Continuous	No

3.1.3 Target Variable

Target	Type	Description	Values	Missing Values
Target	Categorical	Status of the student at the end of the course	0 - Dropout, 1 - Enrolled, 2 - Graduate	No

3.2 Machine Learning Algorithms Applied

Model	Accuracy
Random Forest	0.8628
SVM	0.8515
Neural Network	0.8048
Gradient Boosting	0.8560
XGBoost	0.8470

TABLE 3.1: Model Accuracy

3.3 SHAP Analysis

3.3.1 SHAP Summary Plot

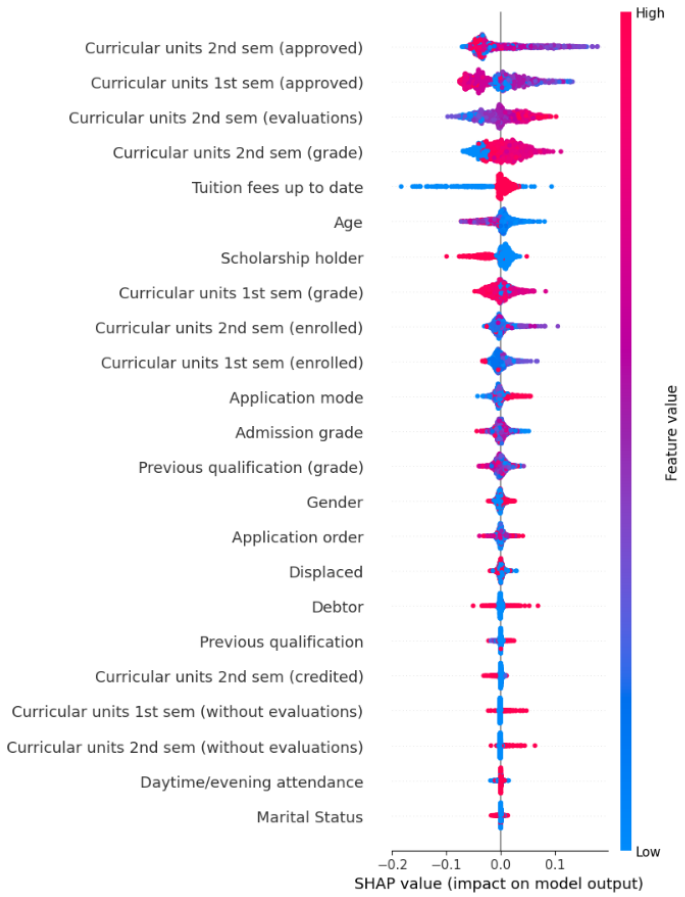

FIGURE 3.1: SHAP Summary Plot

The SHAP Summary plot is a comprehensive visualization that illustrates the impact of various features on the model's output, providing a detailed understanding of how different factors contribute to the predictions. This plot not only ranks the features by their importance but also shows the distribution of the impact each feature has across the entire dataset. The key observations from the SHAP Summary plot are as follows:

- Curricular units 2nd sem (approved) emerges as the feature with the highest impact, clearly indicating that the number of approved curricular units in the second semester plays a crucial role in the dropout prediction. This suggests that students who successfully complete more courses in the second semester are significantly less likely to drop out, highlighting the importance of maintaining academic progress as the academic year progresses.

- Following closely in importance are Curricular units 1st sem (approved) and Curricular units 2nd sem (evaluations), which also exert a strong influence on the model's predictions. The approval of curricular units in the first semester serves as an early indicator of student success, suggesting that those who perform well initially are more likely to continue successfully. Similarly, the evaluations in the second semester are critical, as they reflect the students' ongoing performance and their ability to meet academic requirements.

- Curricular units 2nd sem (grade) and Tuition fees up to date are next in line in terms of their influence on the model's output. The grades obtained in the second semester are pivotal in determining a student's likelihood of persisting in their studies, indicating that strong academic performance later in the year can mitigate the risk of dropout. On the other hand, keeping tuition fees up to date reflects financial stability, which is often a crucial factor in a student's ability to remain enrolled and focused on their studies.

- Additionally, features such as Age, Scholarship holder status, and Curricular units 1st sem (grade) also play substantial roles in the model's predictions. Older students may have different responsibilities and commitments, influencing their academic journey. Being a scholarship holder tends to reduce dropout probability, likely due to the financial support and motivation it provides. Meanwhile, the grades obtained in the first semester contribute to building a solid academic foundation, further reducing the risk of dropout.

- The plot also reveals that features like Marital status, Daytime/evening attendance, and Curricular units 2nd sem (credited) have comparatively lower impacts on the model's predictions. While these factors do contribute to the overall prediction, their influence is less pronounced compared to the primary academic and financial factors. For instance, whether a student attends classes during the day or evening may reflect personal circumstances, but it doesn't have as significant an impact as the core academic performance indicators.

The SHAP Summary plot employs color coding to convey the values of the features, with pink representing higher feature values and blue representing

lower feature values. This color scheme allows for a clear visual interpretation of how feature values correlate with the model's output. For example, higher values of approved curricular units in the second semester (shown in pink) are associated with a decrease in the likelihood of dropout, as indicated by the negative SHAP values. This visual representation makes it easier to identify which feature values contribute positively or negatively to the prediction, enhancing the interpretability of the model.

Overall, the SHAP Summary plot provides a nuanced view of how different features impact the model's predictions, offering both a ranking of feature importance and an understanding of how specific feature values influence the outcome. This comprehensive analysis is invaluable for stakeholders who need to interpret the model's decisions and understand the underlying factors that drive the predictions.

3.3.2 SHAP Bar Plot

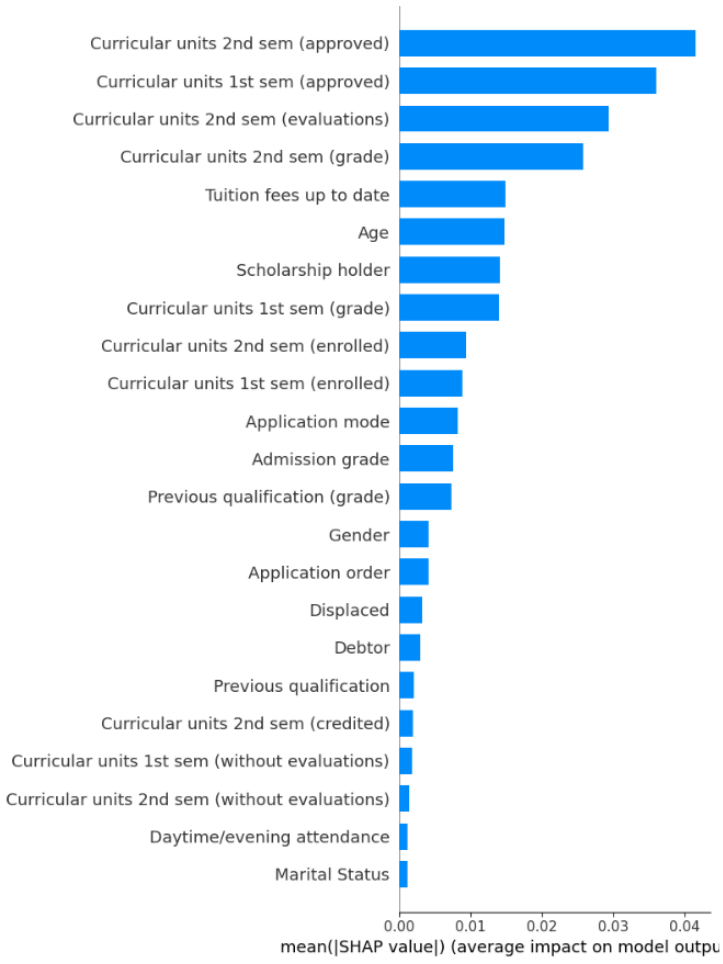

FIGURE 3.2: SHAP Bar Plot

This plot ranks the features according to their mean absolute SHAP values, effectively showcasing their average impact on the model's predictions. By doing so, it provides a clear and comprehensive view of which features are most influential across the entire dataset. This method of ranking is particularly valuable because it aggregates the importance of each feature, giving us an

overall picture of their significance, rather than just focusing on individual instances.

- The ranking presented in the plot reaffirms the insights gleaned from the summary plot, particularly highlighting the importance of Curricular units from the 2nd semester (approved). This feature emerges as the most critical factor influencing the model's predictions, underscoring its role as a strong determinant in whether a student is likely to succeed or face challenges. The approval of curricular units in the second semester appears to be a robust indicator of academic stability and continuity.

- Following this, Curricular units from the 1st semester (approved) is identified as the second most important feature. This emphasizes that academic performance early in the educational journey is also a significant predictor of future success, reinforcing the idea that consistency in academic achievement is crucial. The importance of these first-semester approvals suggests that students who build a solid foundation early on are more likely to maintain momentum in their studies.

- Curricular units from the 2nd semester (evaluations) also ranks highly, indicating that not just the approval of courses but the evaluations themselves are pivotal in predicting outcomes. This feature's position in the ranking highlights the importance of continuous assessment and feedback in shaping a student's academic trajectory. The model seems to capture the nuance that how well students perform in these evaluations can significantly influence their likelihood of persisting in their studies.

- Other significant features that the plot brings to light include Curricular units from the 2nd semester (grade), which suggests that the grades students achieve in their second semester are crucial indicators of their future performance. High grades in these courses likely reflect a strong grasp of the material, which translates into a lower probability of dropout.

- Additionally, Tuition fees up to date is another key feature, indicating that financial stability plays an important role in educational outcomes. Students who are able to keep their tuition fees up to date are likely less burdened by financial stress, which in turn might contribute positively to their academic performance and persistence. This finding points to the broader socio-economic factors that can impact student success, highlighting the intersection of financial and academic considerations.

3.4 LIME Analysis

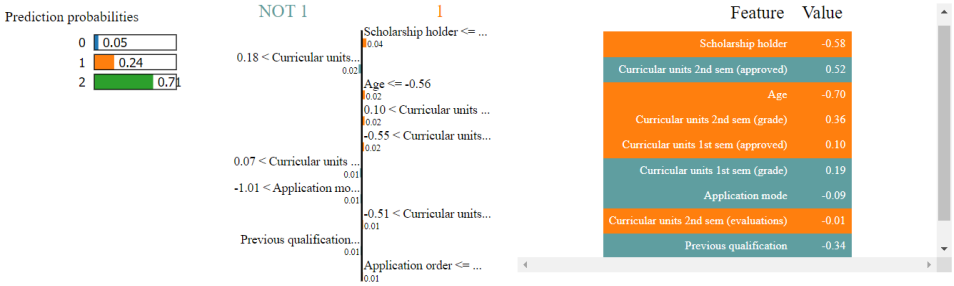

FIGURE 3.3: LIME Plot

The LIME plot provides a detailed explanation for a single prediction, offering an instance-specific explanation of how various features contribute to the model's output. In this particular case, the plot reveals that:

- Scholarship holder status is identified as a critical factor, with a negative influence on the probability of dropout. This means that being a scholarship holder significantly decreases the likelihood of a student dropping out, highlighting the importance of financial support in educational persistence.
- Curricular units from the 2nd semester (approved) are shown to significantly reduce the dropout probability. This feature stands out as a strong indicator of student success, suggesting that passing courses in the second semester plays a crucial role in keeping students on track towards completing their studies.
- Age and Curricular units from the 2nd semester (grade) are also found to reduce the probability of dropout. These features contribute positively to student retention, indicating that older students or those who perform well in their second semester are more likely to continue their education. The impact of grades, in particular, underscores the importance of academic performance in predicting student outcomes.
- Curricular units from the 1st semester (grade) and Application mode are identified as other influential features in the model's prediction. These factors, although not as dominant as the aforementioned ones, still play a significant role in determining the dropout probability, highlighting the multifaceted nature of the dropout risk.
- The LIME plot visually represents the contribution of each feature to the prediction in the right panel. In this visualization, features that negatively influence the dropout probability are marked in orange, while those that positively contribute to it are marked in blue. This color-coded representation helps to quickly discern which features are pushing the prediction towards a higher or lower dropout probability.

3.5 Comparison of SHAP and LIME

In comparing SHAP and LIME, it becomes evident that both methods are effective in identifying key features that influence model predictions, although they differ in focus, methodology, and representation:

- Both SHAP and LIME highlight Curricular units from the 2nd semester (approved) as the most critical feature influencing dropout probability. This agreement between the two methods reinforces the importance of this feature in the predictive model.
- The Scholarship holder status is prominently featured in the LIME plot, where it is shown to be highly influential in reducing dropout probability. However, in the SHAP plots, this feature is less prominently highlighted, suggesting a difference in how each method perceives and ranks the importance of this feature across the dataset.
- Both methods emphasize the significance of grades and evaluations from the first and second semesters, indicating that academic performance is a consistent and important predictor of student success. This convergence on similar features underscores the reliability of both SHAP and LIME in identifying critical factors affecting the model's predictions.
- A key distinction between the two methods lies in their scope of explanation. SHAP provides a global explanation, offering insights into how each feature impacts the model's predictions across the entire dataset. This global perspective allows for a broader understanding of feature importance and interactions. On the other hand, LIME offers a local explanation, focusing on the contribution of features to a single instance's prediction. This localized approach enables a more granular understanding of the specific factors driving an individual prediction, making it particularly useful for case-by-case analysis.

In summary, while SHAP and LIME often identify similar features as important, their differing approaches and representations provide complementary perspectives on model interpretability, offering both global and local insights into the factors driving predictions.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Gallo Giovanni, for his invaluable guidance, support, and encouragement throughout the thesis. His insightful advice and unwavering patience have been instrumental in shaping this thesis, and I am profoundly grateful for the time and effort he dedicated to helping me navigate the complexities of this project. Professor Gallo, your expertise and mentorship have not only enriched my understanding of the subject matter but have also inspired me to strive for excellence in my work. I am incredibly fortunate to have had the opportunity to learn from you, and your contributions to this thesis are immeasurable. Thank you for believing in my abilities and for providing me with the knowledge and confidence to see this research through to completion.