Project Report-German Bank Loan

By Rishab Khatokar

INTRODUCTION

In the dynamic landscape of banking, where risk management is paramount, this project delves into the historical data of leading German Bank. The dataset under scrutiny encapsulates a wealth of information concerning customers who have availed themselves of loans from this institution. With features ranging from checking balance, loan duration, and credit history to employment details, savings balance, and existing loans count, the dataset offers a comprehensive view of customer profiles. The central objective is to construct a robust machine learning model capable of predicting loan default based on this historical data.

As financial institutions grapple with multifaceted challenges, understanding the intricate relationship between customer attributes and the likelihood of loan default becomes instrumental. In this context, we aim to leverage machine learning techniques to not only address the immediate concern of defaulters but also to contribute insights that can fortify the bank's risk management strategies.

This exploration prompts several compelling questions. What are the key determinants of loan default in the context of this German bank? How do various features interplay to influence default risk? Can machine learning models discern patterns within historical data to predict future loan outcomes accurately? By unravelling these questions, we aspire to provide a nuanced understanding of the factors shaping loan dynamics for this financial institution.

METHODS AND MATERIALS

The analysis encompasses both Exploratory Data Analysis(EDA) and the application of various machine learning models. The EDA phase consist of examining the different features/columns and understanding the historic data provided to us.

Exploratory Data Analysis(EDA)

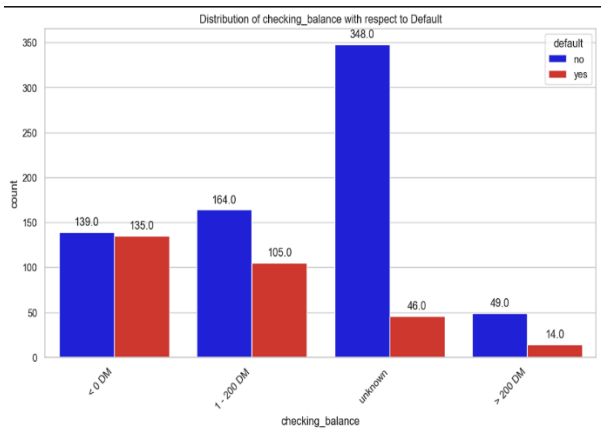
	Attribute	Category Option 1	Category Option 2
0	default	no	yes
1	Percentage	70.000000	30.000000
2	phone	no	yes
3	Percentage	59.600000	40.400000

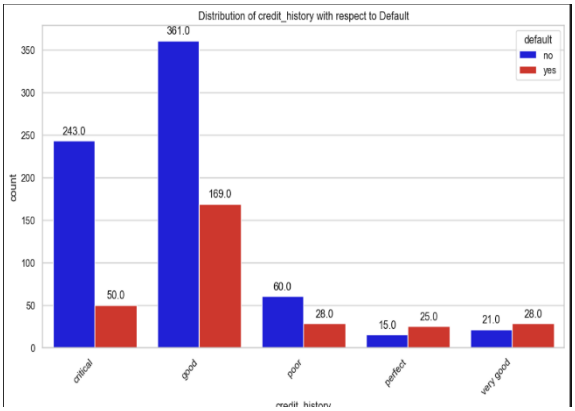
	Attribute	Category Option 1	Category Option 2	Category Option 3
0	housing	other	own	rent
1	Percentage	10.800000	71.300000	17.900000
2	other_credit	bank	none	store
3	Percentage	13.900000	81.400000	4.700000

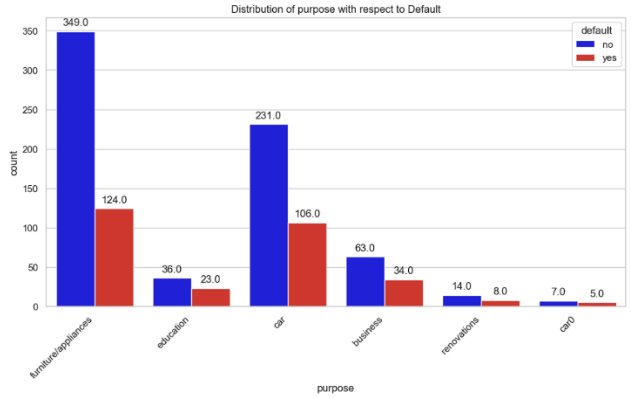
	Attribute	Category Option 1	Category Option 2	Category Option 3	Category Option 4
0	checking_balance	1 - 200 DM	< 0 DM	> 200 DM	unknown
1	Percentage	26.900000	27.400000	6.300000	39.400000
2	job	management	skilled	unemployed	unskilled
3	Percentage	14.800000	63.000000	2.200000	20.000000

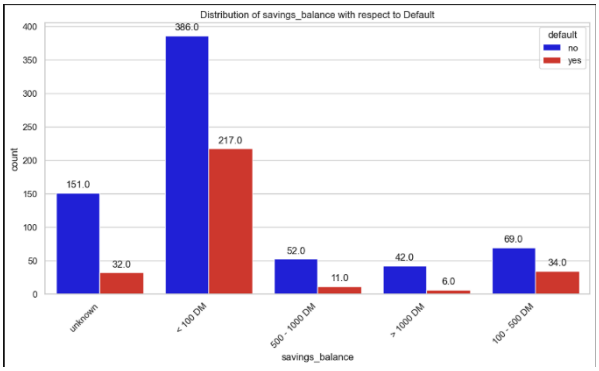
	Attribute	Category Option 1	Category Option 2	Category Option 3	Category Option 4	Category Option 5	Category Option 6
0	purpose	business	car	car0	education	furniture/appliances	renovations
1	Percentage	9.700000	33.700000	1.200000	5.900000	47.300000	2.200000

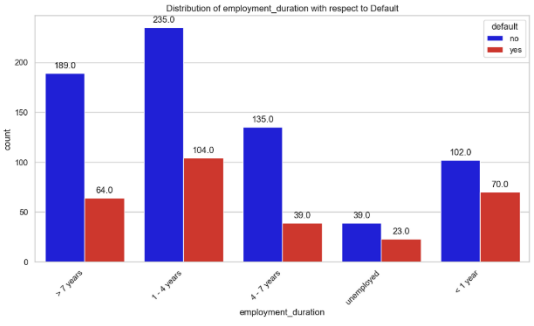
	Attribute	Category Option 1	Category Option 2	Category Option 3	Category Option 4	Category Option 5
0	credit_history	critical	good	perfect	poor	very good
1	Percentage	29.300000	53.000000	4.000000	8.800000	4.900000
2	employment_duration	1 - 4 years	4 - 7 years	< 1 year	> 7 years	unemployed
3	Percentage	33.900000	17.400000	17.200000	25.300000	6.200000
4	savings_balance	100 - 500 DM	500 - 1000 DM	< 100 DM	> 1000 DM	unknown
5	Percentage	10.300000	6.300000	60.300000	4.800000	18.300000

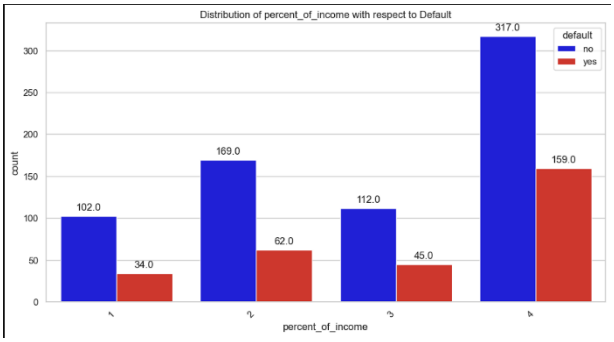


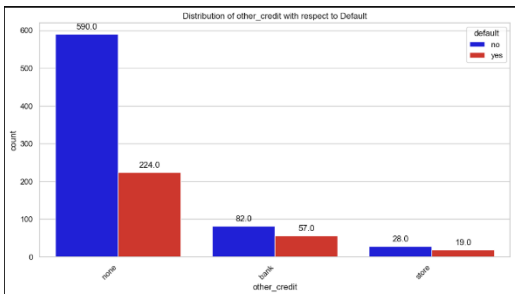


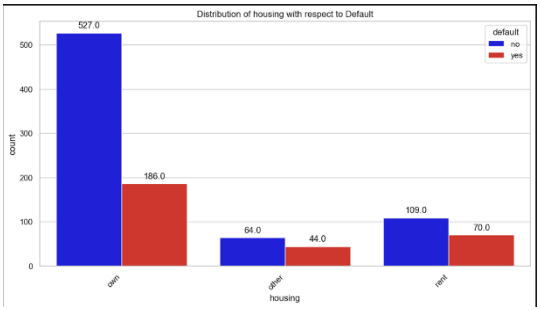


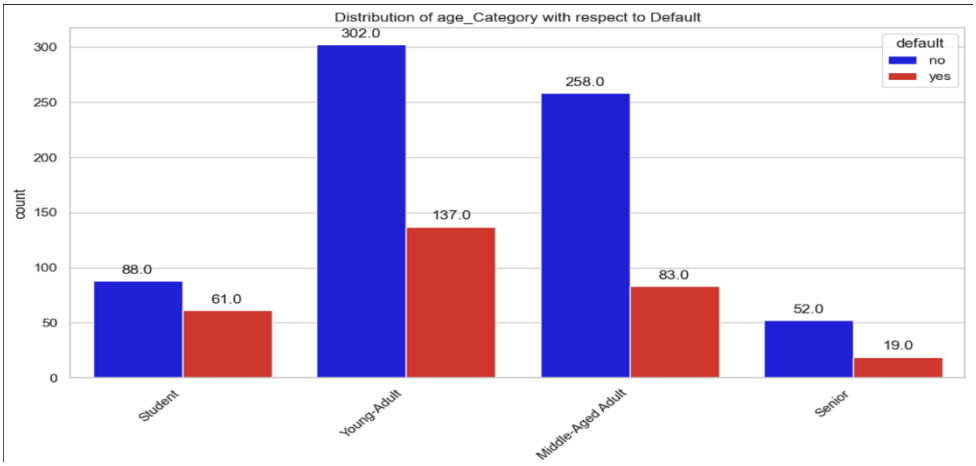


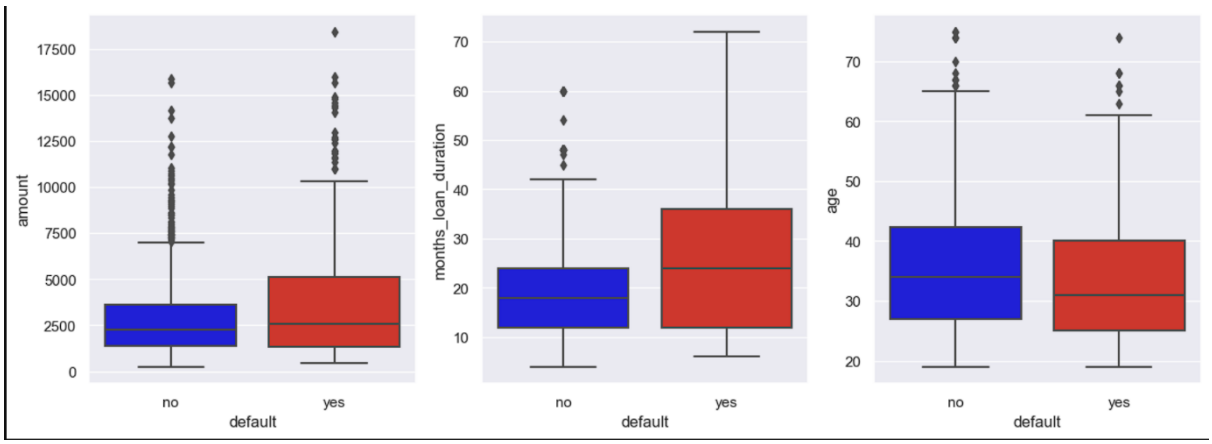






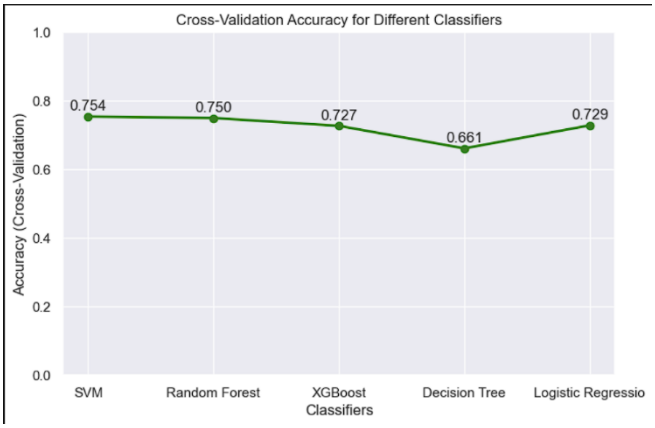






From the above data and visualisations, we can deduce that 30% of the people defaulted in the given data, of those 30% of the people, people who have a credit rating of “good” have defaulted more that that of the “critical” rating. Young adults and middle-aged adults are the age categories in which most of the people have availed the loan compared to the other age categories and most of the young adults defaulting compared to other age categories.

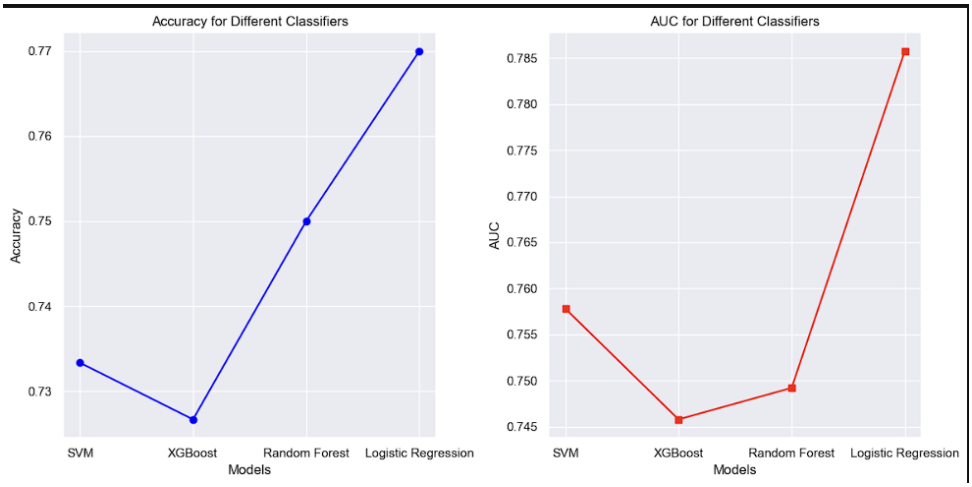
The modelling phase consist of using various algorithms which are Support Vector Machines(SVM), Random Forest, XGBoost, Decision Tree and Logistic Regression classifiers. When each model was cross validated, it was found that the decision tree algorithm was giving less average accuracy compared to the other models, thus it was removed for the final prediction. To compare the different models after the prediction, accuracy and ROC is used.

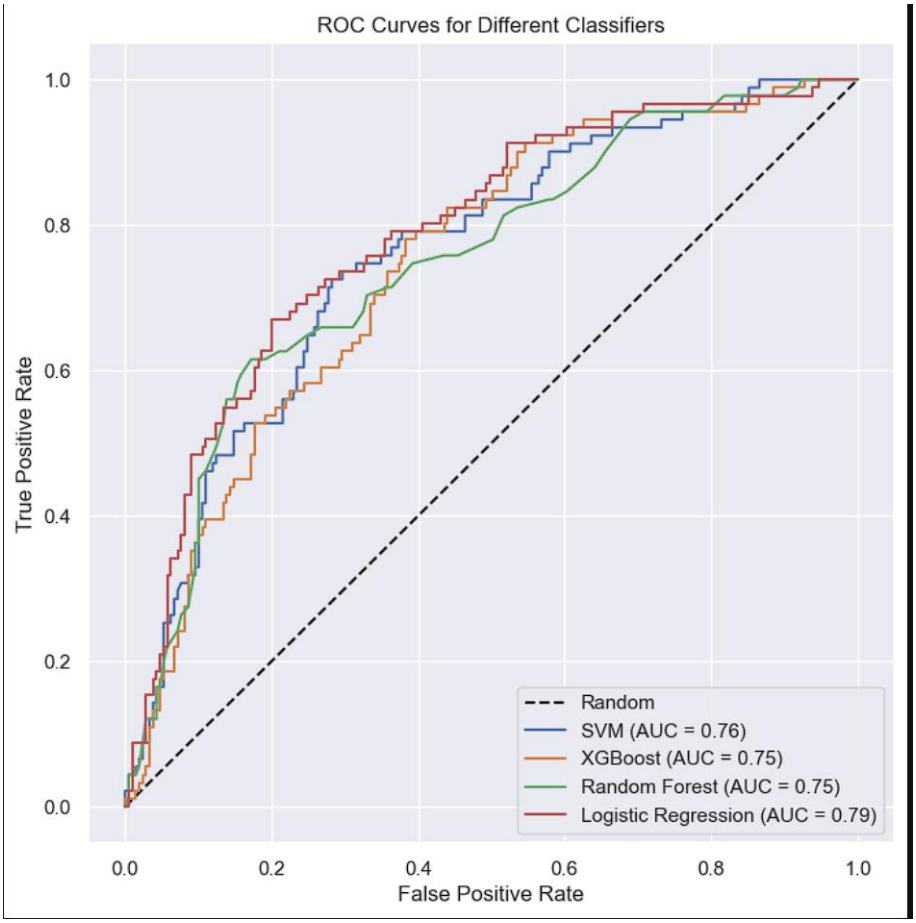


RESULTS

The evaluation of various machine learning models for predicting loan default has yielded insightful results. The table below provides a summary of the model performance metrics, showcasing accuracy and Area Under the Curve (AUC) for each algorithm.

	Model	Accuracy	AUC
0	SVM	0.733333	0.757795
1	XGBoost	0.726667	0.745781
2	Random Forest	0.750000	0.749198
3	Logistic Regression	0.770000	0.785741





The logistic regression model stands out with the highest accuracy of 77%, demonstrating its efficacy in predicting loan default based on the provided historical data. Moreover, the AUC value of 0.7857 underscores the model's robustness in distinguishing between default and non-default instances.

While the Random Forest model boasts a respectable accuracy of 75%, its AUC of 0.7492 indicates a slightly weaker discriminatory power compared to the logistic regression model. SVM and XGBoost, with accuracies of 73.3% and 72.7% respectively, offer competitive performance but fall behind the logistic regression model in terms of overall accuracy.

These results provide valuable insights into the relative strengths of each model in the context of the dataset. The logistic regression model, balancing interpretability and performance, emerges as a promising choice for the German bank's objective of predicting loan default. Further fine-tuning and exploration of ensemble methods could potentially enhance the predictive capabilities of these models.

DISCUSSION

The observed model performances underscore the viability of machine learning in predicting loan default based on historical customer data. The logistic regression model, despite its simplicity, exhibits superior accuracy and AUC, positioning it as a compelling choice for the bank's risk assessment strategy. The emphasis on interpretability in logistic regression aligns with the need for transparency in financial decision-making.

However, it's crucial to consider the specific goals and constraints of the bank when selecting a model. While logistic regression excels in interpretability, more complex models like Random Forest and XGBoost might capture subtle patterns that contribute to improved predictive performance. The SVM model, although competitive, falls slightly behind in accuracy compared to the other algorithms.

Limitations of the study include the reliance on historical data, potentially overlooking real-time economic shifts. Future research could explore dynamic models that adapt to evolving economic conditions and incorporate additional external factors for a more comprehensive risk assessment.

CONCLUSIONS

In conclusion, the evaluation of machine learning models for predicting loan default reveals the logistic regression model as a frontrunner in accuracy and AUC. The practical balance it strikes between interpretability and performance makes it a pragmatic choice for the German bank. As financial institutions navigate the intricate landscape of risk management, these findings provide a foundation for informed decision-making and underscore the potential of machine learning in enhancing predictive analytics for loan default.