**WHITE PAPER/REPORT**

**Predicting Loan Default Risk Using Customer Financial Profiles**

**By Serge Nane**

**DSC680 Applied Data Science**

**Fall 2025**

**Table of contents**

1. Introduction .................................................... 3

2. Dataset and Methodology .......................................... 3

3. Results .......................................................... 5

4. Discussion ....................................................... 6

5. Ethical Considerations ........................................... 7

6. Conclusion and Future Work ....................................... 7

7. References .......................................................8

8. Appendix .........................................................8

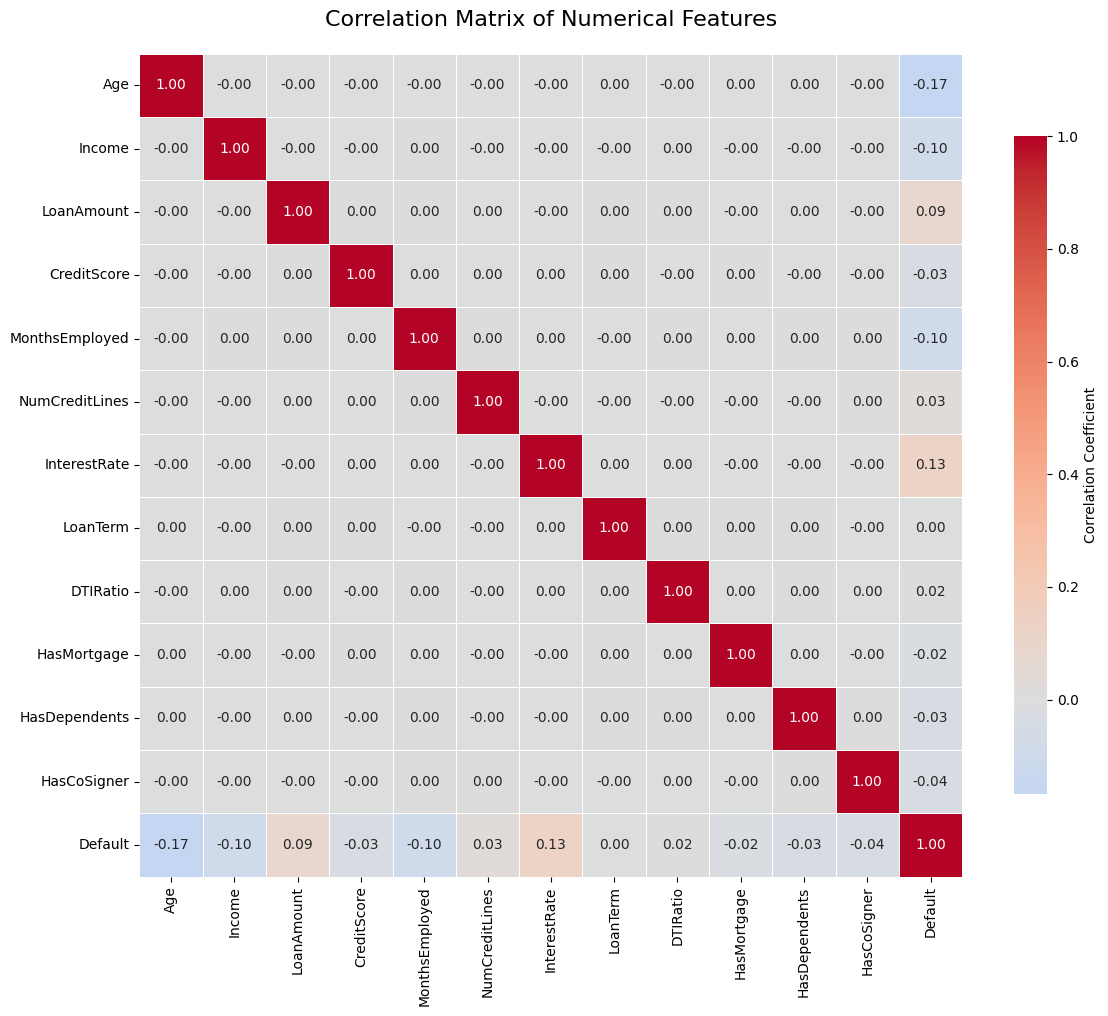
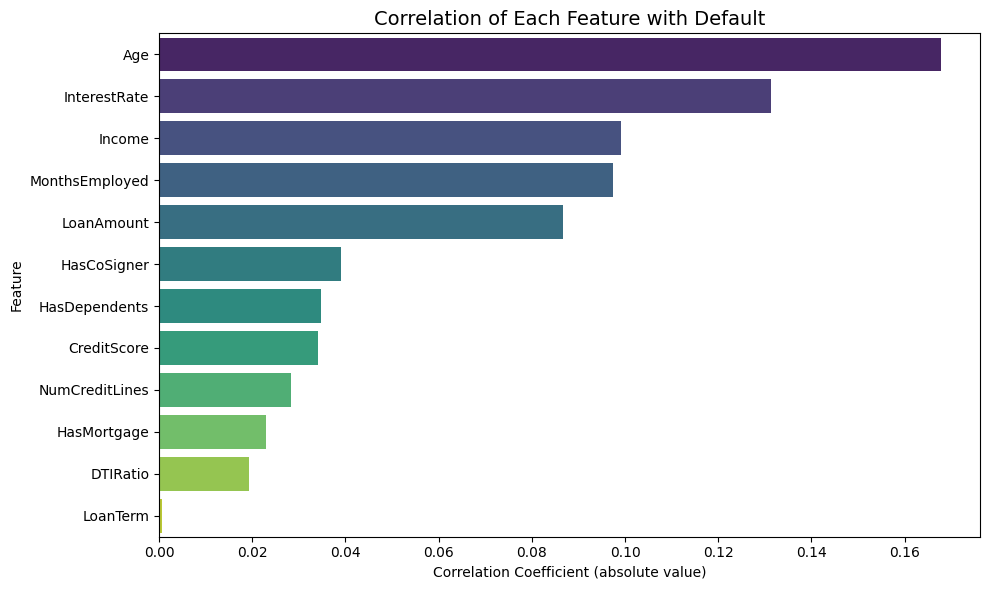
9. 10 Anticipated Audience Questions ................................ 9

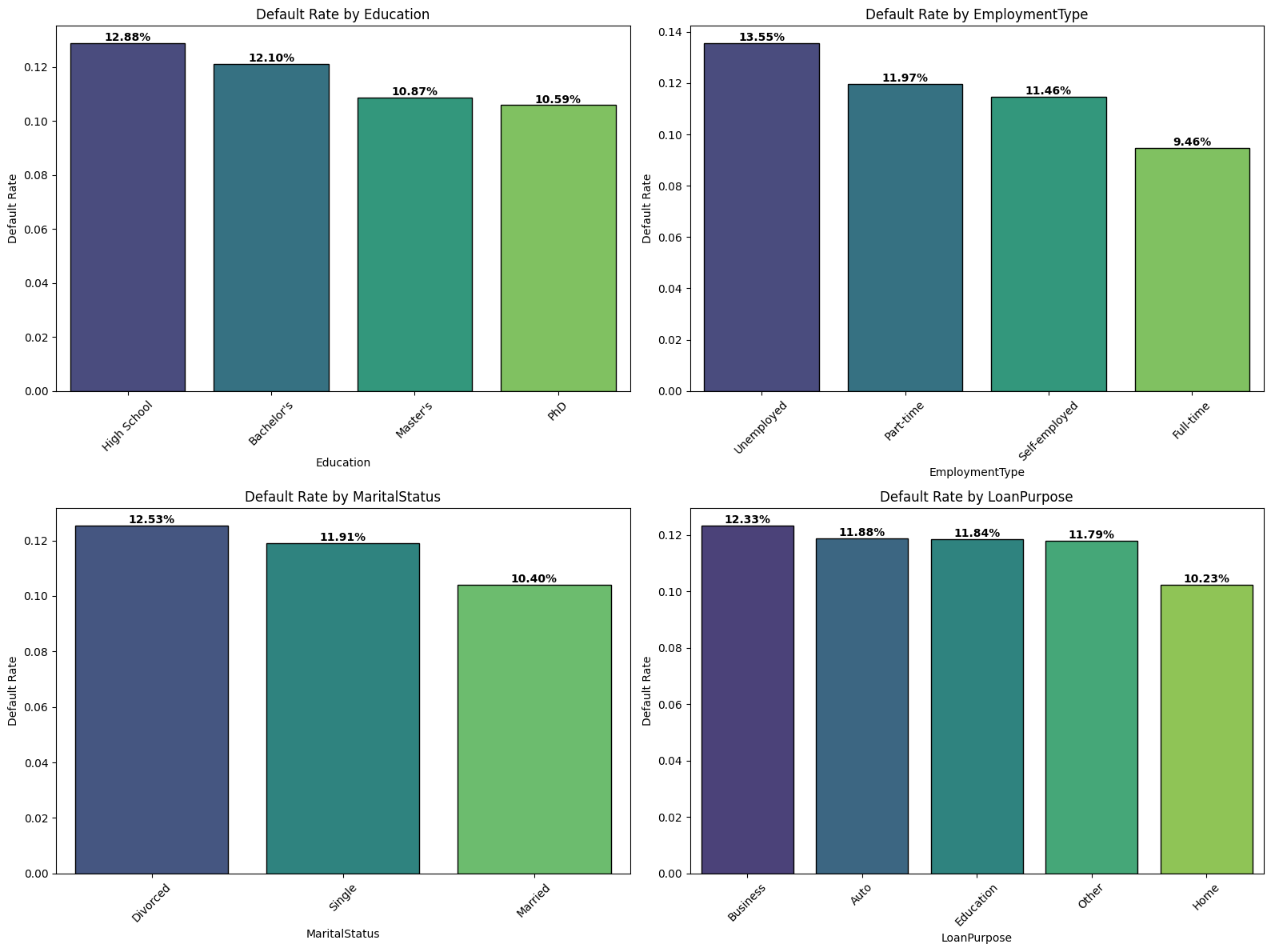
**1. Introduction**

Loan default represents a significant financial risk to lending institutions, directly impacting profitability and economic stability. Traditional credit decision methods often rely on limited financial indicators such as credit scores and income levels, which may not fully capture the complex relationships between multiple borrower attributes that influence repayment behavior. This project applies machine learning to customer demographic and financial data to predict the probability of loan default. We evaluated and compared Logistic Regression and Random Forest models to determine the most effective approach for risk assessment and identify which borrower characteristics most strongly influence default probability.

**2. Dataset and Methodology**

The LendingClub Loan Default Prediction Dataset from Kaggle was used, consisting of 255,347 loan records with 17 features and 1 target variable (Default). The dataset includes borrower information such as age, income, loan amount, credit score, employment length, interest rate, and debt-to-income ratio. Preprocessing included removing the unique loan identifier (LoanID), converting binary "Yes"/"No" fields to numerical values, and addressing class imbalance (11.61% default rate) using SMOTE oversampling on the training data. Exploratory data analysis examined variable distributions, correlations, and default rates across categorical groups. Models (Logistic Regression, Random Forest) were trained on a stratified 80/20 split. Numerical features were standardized, and categorical features were one-hot encoded using a ColumnTransformer. Model evaluation was based on **Accuracy, Precision, Recall, F1-score,** and **ROC-AUC.**

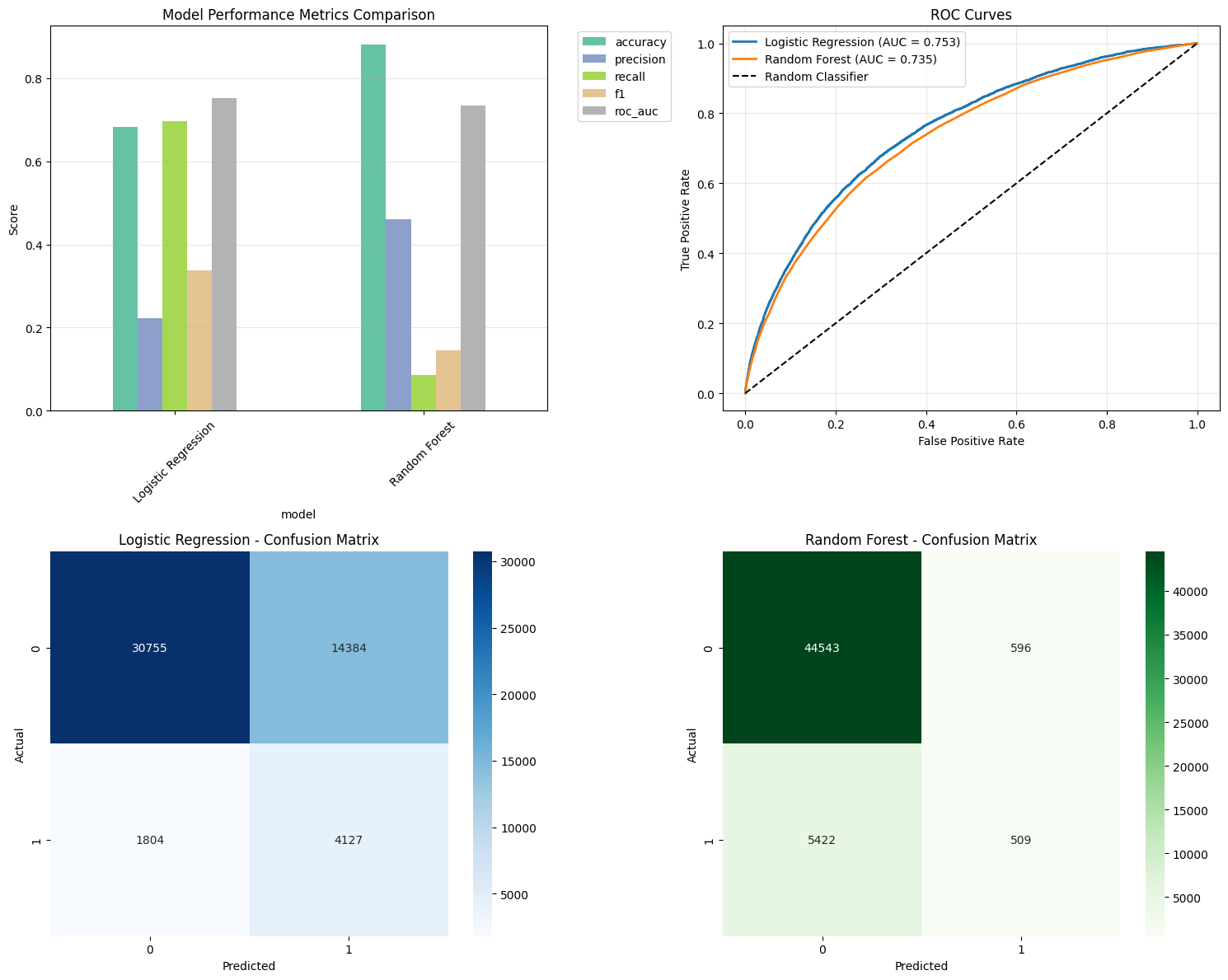
****

****

**3. Results**

The model evaluation revealed a critical business trade-off. While **Random Forest** achieved higher overall **Accuracy (0.88)**, **Logistic Regression** demonstrated superior performance on metrics more relevant to risk detection, including **Recall (0.70)** and **ROC-AUC (0.753)**.

Logistic Regression correctly identified 70% of actual defaulters, making it significantly more effective at detecting risky borrowers compared to Random Forest, which identified only 9% of defaulters. Feature importance analysis consistently identified **Age, Interest Rate, Months Employed, and Income** as the most influential predictors across both models.



**4. Discussion**

These results demonstrate that for high-stakes financial decisions like loan approval, model selection must prioritize business objectives over pure accuracy metrics. Logistic Regression emerges as the more suitable model for this context because its high recall directly addresses the primary business need: minimizing financial losses from undetected defaulters. While Logistic Regression generates more false positives (potentially leading to good customers being flagged for additional review), this is preferable to the alternative of missing actual defaulters, which represents direct financial loss. The model's interpretability provides clear, actionable insights for financial institutions, showing that younger borrowers, those with higher interest rates, unemployed individuals, and applicants with shorter employment history present significantly higher default risk.

**5. Ethical Considerations**

* **Data Privacy:** Although the dataset is anonymized, handling customer financial data requires strict adherence to data privacy regulations and ethical guidelines for financial information.
* **Bias and Fairness:** The models must be carefully monitored to ensure predictions do not disproportionately negatively impact specific demographic groups based on age, employment type, or other characteristics that could correlate with protected attributes.
* **Interpretability:** The Logistic Regression model offers superior interpretability through clear coefficient values, which is crucial for regulatory compliance and for building trust with both lenders and borrowers. The "black-box" nature of ensemble methods like Random Forest makes their decisions harder to justify in financial contexts.
* **Responsible Use:** This model should be deployed as a decision-support tool to augment the expertise of human loan officers, not to replace them. Final lending decisions should involve a holistic review that considers factors beyond the model's binary output.

**6. Conclusion and Future Work**

Logistic Regression is recommended as the most practical and business-aligned model for predicting loan default due to its high recall and interpretability. The analysis confirms that borrower age, interest rate, employment stability, and income are the strongest predictors of default risk. Future work should focus on threshold tuning to optimize the precision-recall trade-off, testing gradient boosting algorithms like XGBoost, and incorporating external economic factors to improve model robustness and predictive power across different economic cycles.

**7. References**

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Kaggle. (n.d.). Loan Default Prediction Dataset. Retrieved from <https://www.kaggle.com/datasets/>

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569.

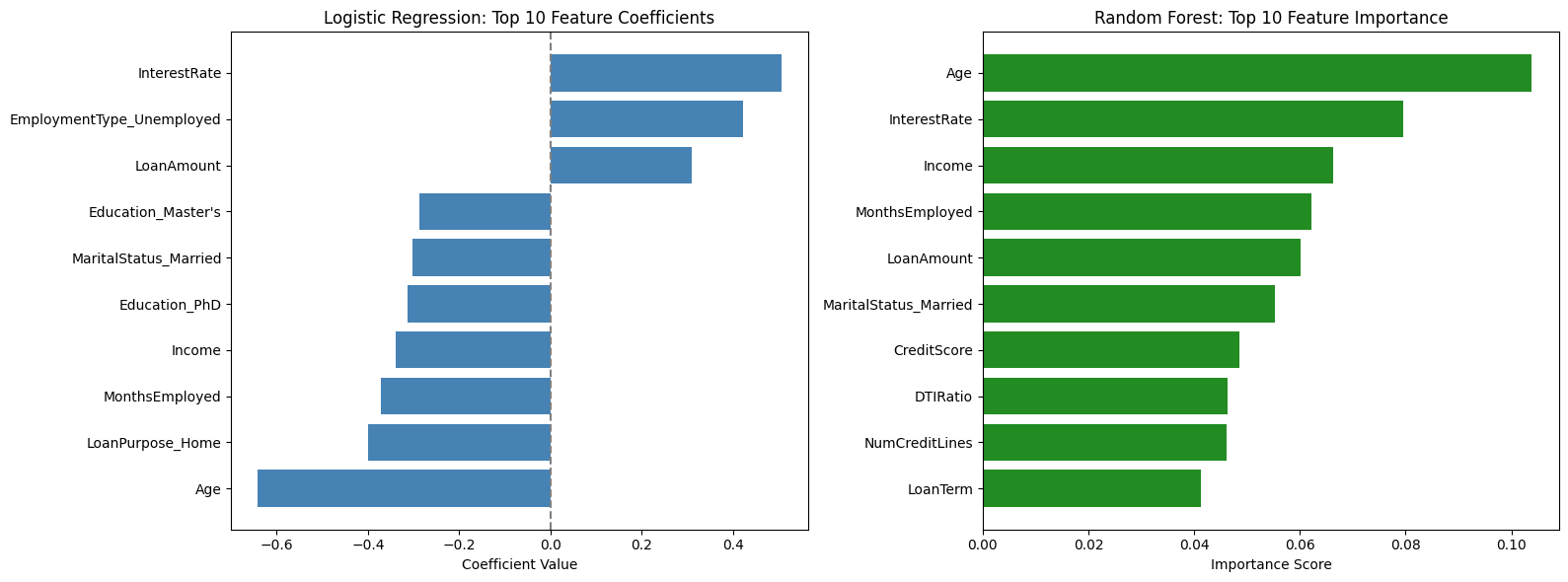
**8. Appendix**

Full code available in the attached Jupyter Notebook and PDF.  

The detailed model performance metrics table for both models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **ROC-AUC** |
| **Logistic Regression** | 0.68 | 0.22 | 0.70 | 0.753 |
| **Random Forest** | 0.88 | 0.46 | 0.09 | 0.735 |

**Confusion Matrix details**:

* **Logistic Regression:** 30,746 True Negatives, 14,393 False Positives, 1,806 False Negatives, 4,125 True Positives
* **Random Forest:** 44,544 True Negatives, 595 False Positives, 5,415 False Negatives, 516 True Positives

**9. 10 Anticipated Audience Questions**

1. Why was the LendingClub dataset chosen for this analysis?
2. How did you address the significant class imbalance in the dataset?
3. Why did you choose to compare Logistic Regression and Random Forest specifically?
4. Why is Logistic Regression considered better despite its lower accuracy?
5. What do the feature importance results tell us about the primary drivers of loan default?
6. How could this model be integrated into a real-world loan approval process?
7. What are the business implications of the False Positives generated by the Logistic Regression model?
8. What ethical risks are associated with using models like this for lending decisions?
9. How would you improve the model's performance in future iterations?
10. What steps are necessary to validate this model before deployment in a financial institution?