Capstone Project

Credit Risk Assessment Final Report

1. Define the Problem Statement

This project develops a machine learning model to predict a loan applicant's Credit Risk Score using diverse demographic, financial, and loan-related features. The objective is to improve the accuracy, efficiency, and fairness of Credit Risk Assessment compared to conventional evaluation methods. The model also addresses key challenges such as feature complexity, interdependence, and the need for interpretability to ensure transparent and data-driven lending decisions.

2. Model Outcomes or Predictions

The proposed solution employs a supervised machine learning approach to predict the Credit Risk Score[1] of loan applicants — a continuous measure of credit quality — making regression the appropriate modelling technique. The model incorporates regularisation methods (Ridge and Lasso regression) to enhance stability and mitigate multicollinearity. Polynomial feature transformations are applied to numerical features to capture higher-order relationships, while One-Hot Encoding is used to convert categorical attributes into a suitable numerical format. To ensure robust and reliable performance, GridSearchCV with cross-validation is utilised for systematic hyperparameter tuning and optimal predictive accuracy.

3. Data Acquisition

Data[2] for this project is sourced from Kaggle, a reputable platform providing publicly available and well-structured datasets. Given logistical and privacy constraints, independent collection of actual credit applicant data is not feasible. The chosen dataset includes a comprehensive set of variables — such as demographic information, employment status, income, debt ratios, credit history, and loan characteristics — that reflect real-world credit assessments. **Exploratory Data Analysis (EDA)** and visualisations, including count plots for categorical features, and histograms and correlation heatmaps for numerical features, reveal meaningful relationships with the target Credit Risk Score, confirming the dataset's suitability for developing an effective predictive model.

[1] A lower score implies a higher credit quality.
[2] https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?select=Loan.csv

The dataset includes the following:

- **Categorical Variables:**
  1. **ApplicationDate**: Loan application date
  2. **EmploymentStatus**: Job situation
  3. **EducationLevel**: Highest education attained
  4. **MaritalStatus**: Applicant's marital state
  5. **HomeOwnershipStatus**: Homeownership type
  6. **LoanPurpose**: Reason for loan

- **Numerical Variables:**
  1. **Age**: Applicant's age
  2. **AnnualIncome**: Yearly income
  3. **CreditScore**: Creditworthiness score
  4. **Experience**: Work experience
  5. **LoanAmount**: Requested loan size
  6. **LoanDuration**: Loan repayment period
  7. **NumberOfDependents**: Number of dependents
  8. **MonthlyDebtPayments**: Monthly debt obligations
  9. **CreditCardUtilizationRate**: Credit card usage percentage
  10. **NumberOfOpenCreditLines**: Active credit lines
  11. **NumberOfCreditInquiries**: Credit checks count
  12. **DebtToIncomeRatio**: Debt to income proportion
  13. **BankruptcyHistory**: Bankruptcy records
  14. **PreviousLoanDefaults**: Prior loan defaults
  15. **PaymentHistory**: Past payment behaviour
  16. **LengthOfCreditHistory**: Credit history duration
  17. **SavingsAccountBalance**: Savings account amount
  18. **CheckingAccountBalance**: Checking account funds
  19. **TotalAssets**: Total owned assets
  20. **TotalLiabilities**: Total owed debts
  21. **MonthlyIncome**: Income per month
  22. **UtilityBillsPaymentHistory**: Utility payment record
  23. **JobTenure**: Job duration
  24. **NetWorth**: Total financial worth
  25. **BaseInterestRate**: Starting interest rate
  26. **InterestRate**: Applied interest rate
  27. **MonthlyLoanPayment**: Monthly loan payment
  28. **TotalDebtToIncomeRatio**: Total debt against income
  29. **LoanApproved**: Loan approval status

- **Target Variable (Numerical):**

  ✓ **RiskScore**: Credit Risk Assessment score

Figure 1: Histogram for 30 Numerical Variables (including target)

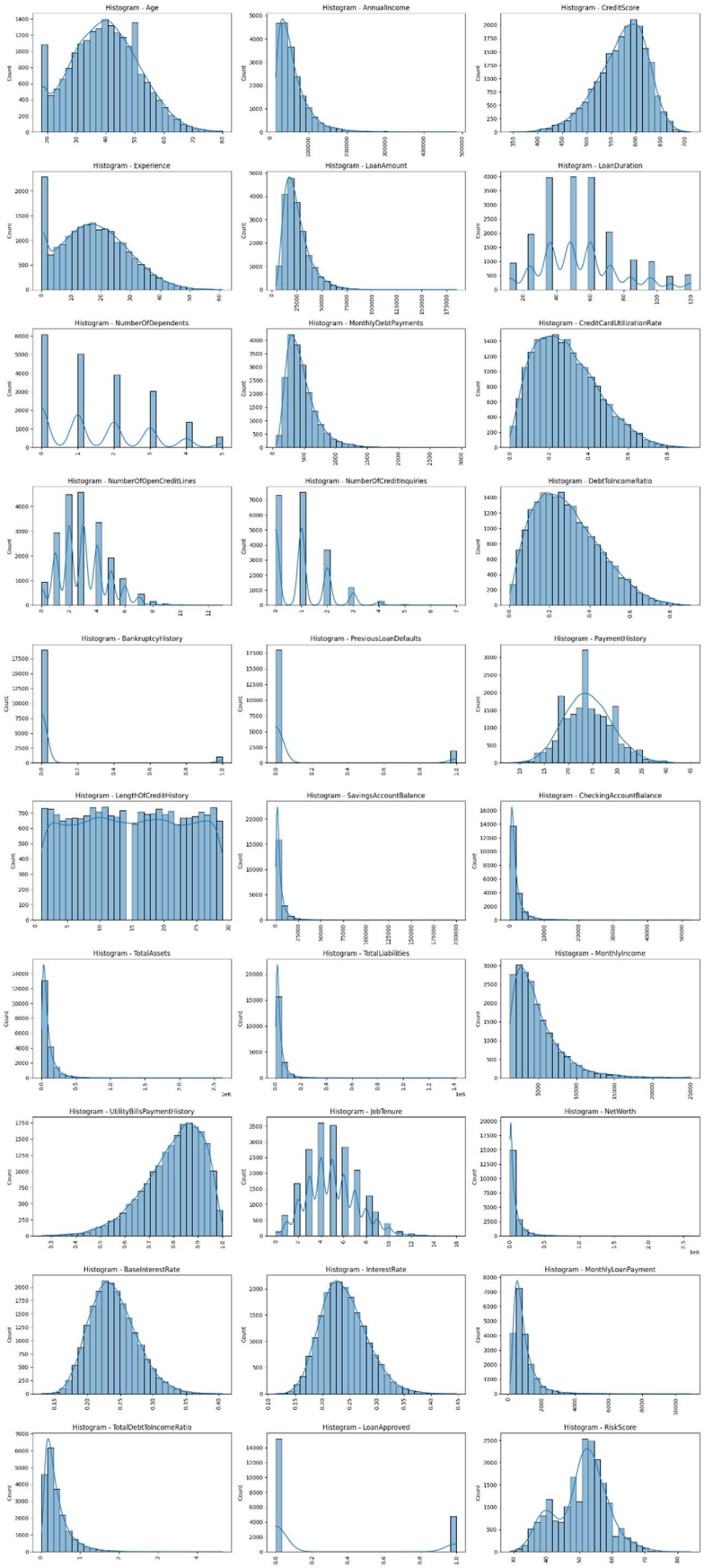


Figure 2: Correlation Heatmap for 30 Numerical Variables (including target)

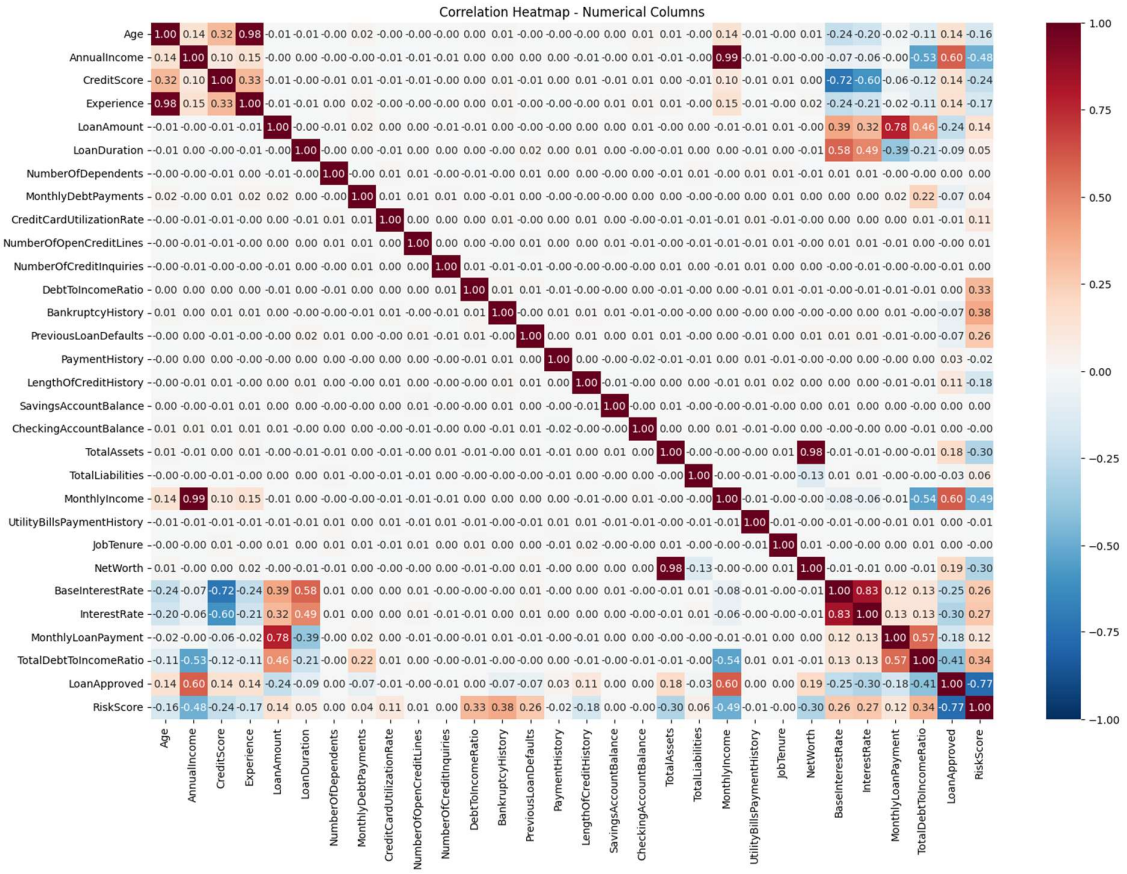
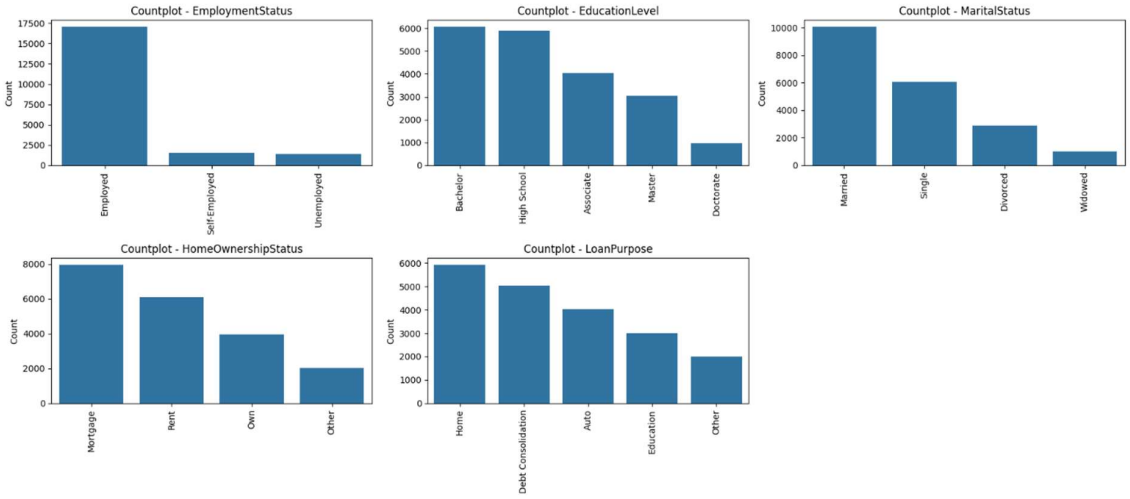


Figure 3: Count plot for 5 Categorical Variables (excluding **ApplicationDate**)



4. Data Preprocessing / Preparation

Data Cleaning and Feature Selection

The dataset used in this study was complete and free of missing values, requiring no imputation. The **ApplicationDate** feature was removed, as temporal information was not considered predictive of the target variable, **RiskScore**. Categorical variables were examined for sparsely populated categories, and none were identified, eliminating the need for category consolidation prior to One-Hot Encoding. Numerical features were reviewed for outliers or invalid entries, and none were observed.

Correlation analysis among numerical variables revealed redundancy, leading to the removal of the following features to reduce multicollinearity: **Experience, MonthlyIncome, MonthlyLoanPayment, NetWorth, BaseInterestRate**.

Additionally, features exhibiting negligible correlation with all other variables — including **NumberOfDependents, NumberOfOpenCreditLines, NumberOfCreditInquiries, PaymentHistory, SavingsAccountBalance, CheckingAccountBalance, UtilityBillsPaymentHistory, JobTenure** — were excluded.

The resulting dataset comprised:

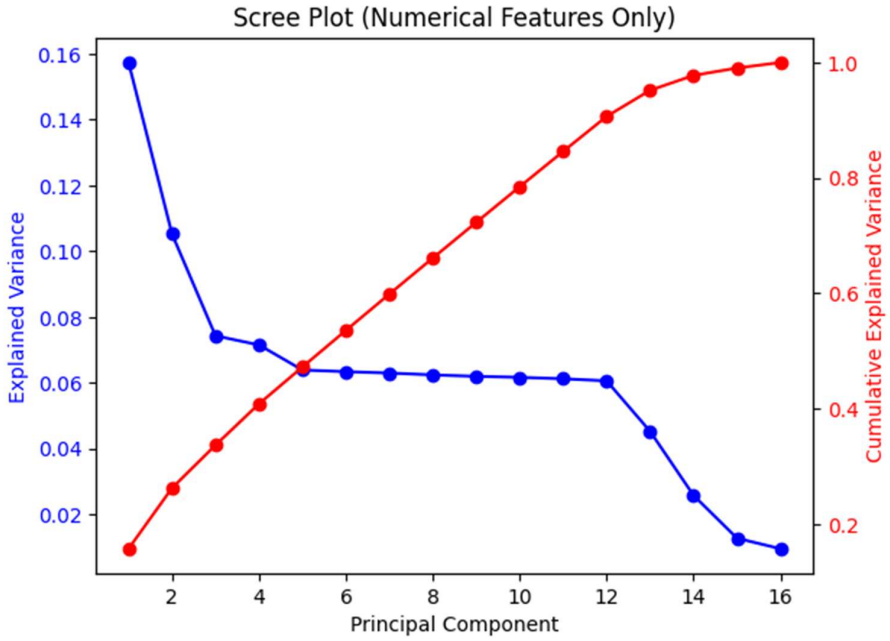
- **Target Variable:**
  - ✓ **RiskScore**

- **Numerical Features:**
  1. **Age**
  2. **AnnualIncome**
  3. **CreditScore**
  4. **LoanAmount**
  5. **LoanDuration**
  6. **MonthlyDebtPayments**
  7. **CreditCardUtilizationRate**
  8. **DebtToIncomeRatio**
  9. **BankruptcyHistory**
  10. **PreviousLoanDefaults**
  11. **LengthOfCreditHistory**
  12. **TotalAssets**
  13. **TotalLiabilities**
  14. **InterestRate**
  15. **TotalDebtToIncomeRatio**
  16. **LoanApproved**

- **Categorical Features:**
    1. **EmploymentStatus**
    2. **EducationLevel**
    3. **MaritalStatus**
    4. **HomeOwnershipStatus**
    5. **LoanPurpose**

Principal Component Analysis (PCA) was applied to assess the potential for dimensionality reduction among the numerical features. The scree plot indicated that 12 components were necessary to capture at least 90% of the variance, suggesting minimal reduction in dimensionality. Considering the limited reduction and the loss of feature interpretability, PCA was not employed in subsequent modelling.

Figure 4: Scree plot for Numerical Features



Data Splitting

The preprocessed dataset was partitioned into training and testing sets using an 80/20 split, ensuring that sufficient data remained for model development while reserving a representative subset for evaluating model performance.

Feature Engineering and Encoding

To enhance the predictive capability of the model, **Polynomial Feature Expansion** was applied to the numerical variables to capture potential non-linear relationships. **One-Hot Encoding** was used for categorical variables to transform them into a numerical format suitable for regression modeling. These preprocessing steps produced a clean, consistent, and analytically robust dataset, providing a strong foundation for developing an accurate and interpretable predictive model for the Credit Risk Score.

Figure 5: Scatter plot of Target vs 16 selected Numerical Features (coloured by 5 Categorical Features)

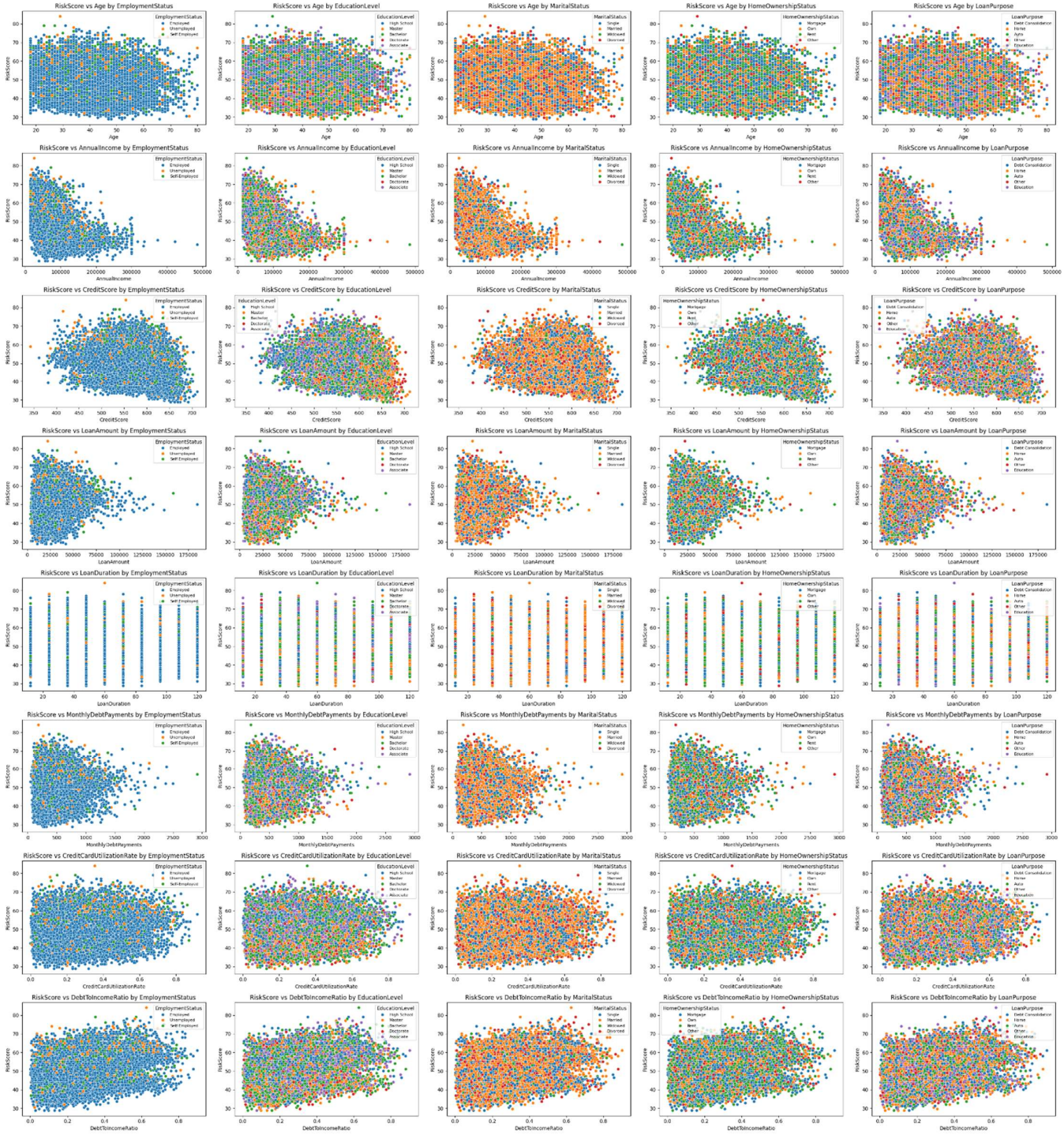
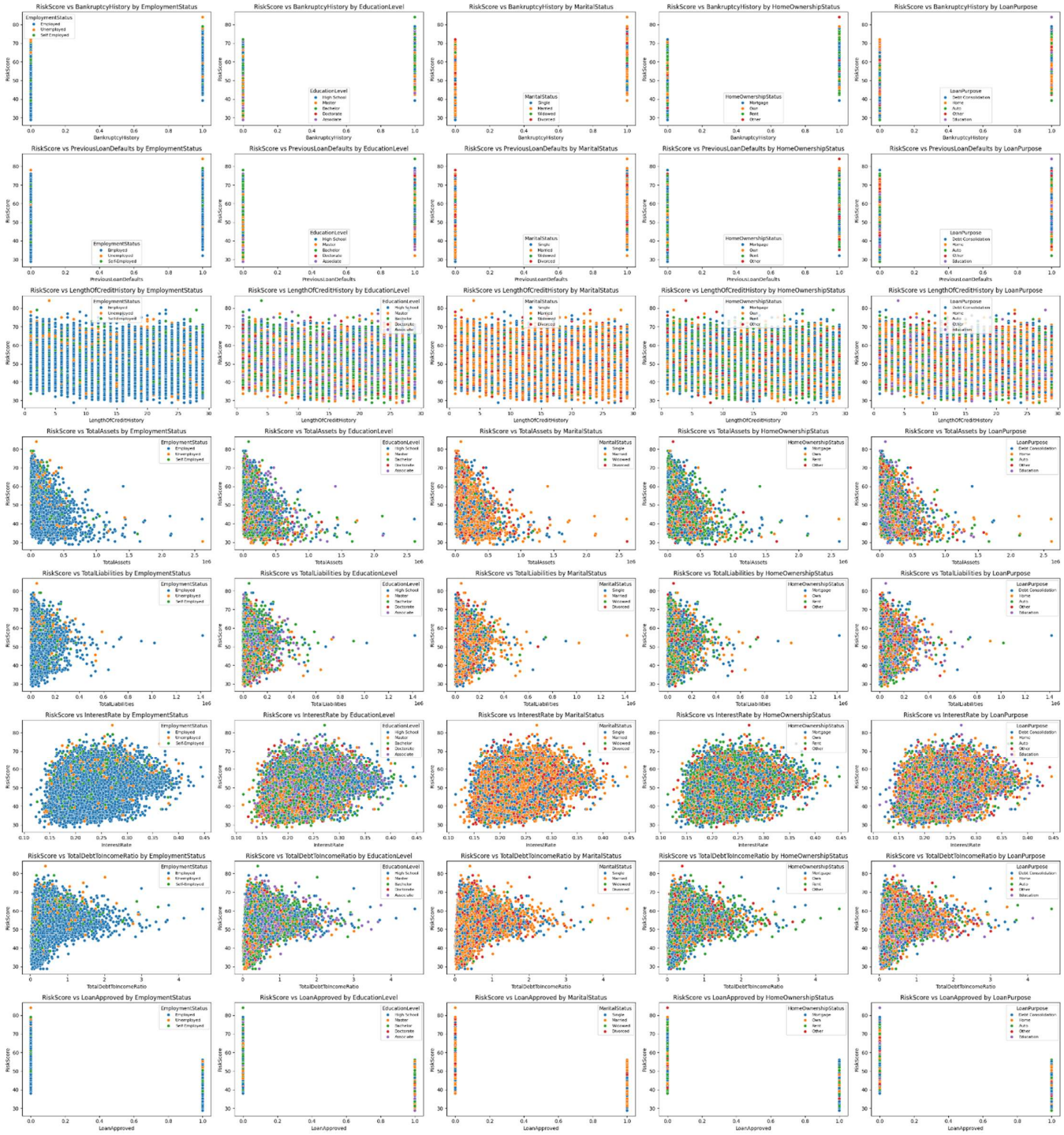


Figure 5 (continued)



5. Modelling

Key supervised machine learning principles were applied to develop a regression model that predicts an applicant's Credit Risk Score based on demographic, financial, and loan-related features. The problem was formulated as a regression task, requiring methods that produce continuous numerical predictions rather than categorical classifications.

In selecting the modelling techniques, a deliberate focus was placed on model simplicity, explainability, and transparency, given the importance of interpretability in financial risk assessments. Simple, well-understood models allow credit analysts and regulatory stakeholders to more easily trace how input features influence predictions, making them preferable in sensitive decision-making contexts such as Credit Scoring.

At the same time, the potential of more complex models, such as Neural Networks, was explored to achieve higher predictive precision. While these models are less inherently interpretable, they can capture intricate patterns and nonlinear relationships in the data, offering superior accuracy for Credit Risk prediction. This dual approach balances the trade-off between predictive performance and model explainability, allowing for highly accurate predictions while retaining simpler models for transparency when needed.

Five models were constructed to explore these trade-offs:

1. **Linear Regression** – established a baseline to assess the degree of polynomial features that best captured the linear relationship with the target variable.

2. **Ridge Regression (L2 regularisation)** – introduced to manage multicollinearity and reduce overfitting by penalising excessively large coefficient values.

3. **Lasso Regression (L1 regularisation)** – employed for both regularisation and feature selection by shrinking less influential coefficients to zero.

4. **Decision Tree Regressor** – implemented to capture nonlinear relationships and complex feature interactions without the need for explicit feature transformations.

5. **Neural Network Regressor** – included as a flexible non-linear model capable of learning complex patterns in the data, allowing comparison with simpler models in terms of predictive performance.

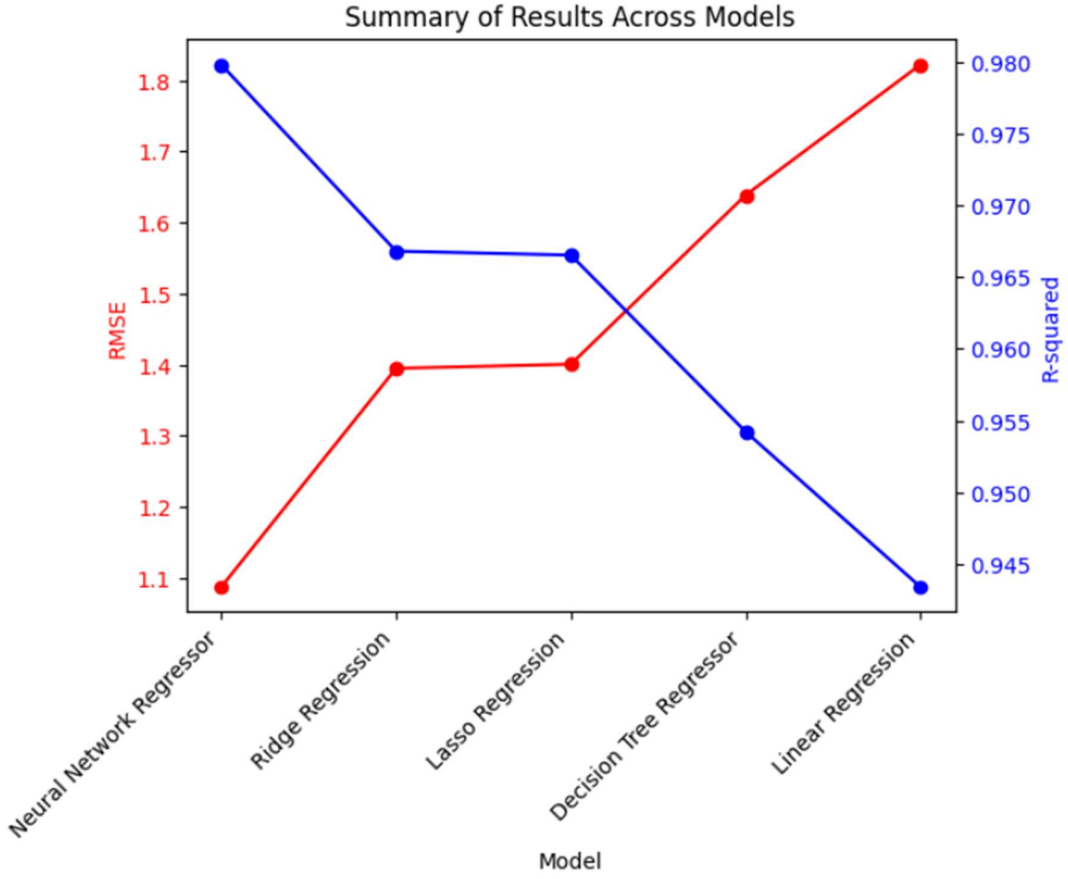
The data was divided using an **80/20 train-test split**, following standard practice to prevent data leakage and to fairly evaluate model performance on unseen data. For model optimisation, **GridSearchCV** with **5-fold Cross-Validation** was applied to identify the optimal hyperparameter values for Ridge and Lasso regression models. This process reinforced the concept of model validation, ensuring that performance metrics generalised well beyond the training data. The inclusion of the Neural Network model also provides insight into potential gains from more complex models, balancing accuracy with interpretability considerations.

6. Model Evaluation

Evaluation metrics were selected to demonstrate core regression evaluation concepts:

- **Root Mean Squared Error (RMSE)** was used to measure the model's prediction error magnitude in the same units as the target variable.
- **R-squared (R^2)** quantified how much of the variance in **RiskScore** was explained by the model, indicating its overall explanatory power.

Figure 6: Model Evaluation Summary



Model comparison demonstrated the trade-off between model complexity, predictive power, and interpretability:

- **Linear Regression** achieved a reasonable fit (**Test RMSE = 1.8220** and **Test R^2 = 0.9434**), serving as a baseline but underperforming compared to regularised models.
- **Ridge Regression (α = 0.1)** performed well **(Test RMSE = 1.3948** and **Test R^2 = 0.9668)** offering strong interpretability and stable generalisation, though not the highest predictive accuracy.
- **Lasso Regression (α = 0.001)** performed comparably **(Test RMSE = 1.4006** and **Test R^2 = 0.9665)** to Ridge Regression but is slightly less stable due to coefficient shrinkage.
- **Decision Tree Regressor** achieved perfect training accuracy but lower test performance **(Test RMSE = 1.6384** and **Test R^2 = 0.9542)**, illustrating the concept of overfitting.
- **Neural Network Regressor** achieved the highest predictive accuracy (**Test RMSE = 1.0874** and **Test R^2 = 0.9798**), demonstrating superior predictive power, though with increased complexity and reduced interpretability compared to the simpler models.

These results illustrate key learning outcomes:

- **Regularisation** improves model robustness by controlling coefficient magnitude, helping to prevent overfitting.
- **Cross-validation** ensures reliable performance evaluation across data subsets, ensuring generalisability.
- **Bias–variance trade-off** explains why the Ridge model achieved strong performance compared to both unregularised Linear Regression and the overfitted Decision Tree.
- **Model complexity vs predictive power** demonstrates that the Neural Network Regressor delivers the highest predictive accuracy, capturing complex patterns in the data and enabling more precise Credit Risk predictions for better-informed lending decisions.

Based on the model evaluation, the **Neural Network Regressor** was selected as the primary model due to its **highest predictive accuracy** among all candidates.

In terms of interpretability, **Ridge Regression** remains a strong alternative when explainability is critical, offering slightly lower accuracy but clear and actionable insights for stakeholders. Key predictors identified include **CreditScore**, **LengthOfCreditHistory²**, and **DebtToIncomeRatio** emerging for higher Credit Risk, and **CreditScore²**, **TotalAssets**, and **LengthOfCreditHistory** for lower Credit Risk.

7. **Deployment**

Email to stakeholders on the business outcome achieved by the Credit Risk Score prediction model:

Dear Stakeholders,

I am pleased to share an update on the Credit Risk Score Prediction Model developed as part of our analytical initiative to enhance the Credit Assessment process for loans. The model is designed to help us estimate an applicant's Credit Risk Score more accurately and consistently by analysing key financial and personal factors that influence repayment behaviour.

Implementing this model provides several important business benefits:

1. **Faster and more consistent decisions:** The model can assess risk in seconds, reducing manual review time and ensuring that every applicant is evaluated using the same criteria.

2. **Improved risk management:** By identifying higher-risk applicants more reliably, we can minimise potential loan defaults and protect the organisation from avoidable financial losses.

3. **Better customer targeting:** The model helps us recognise applicants with strong credit potential, allowing us to approve quality loans more confidently and offer more competitive products to low-risk customers.

4. **High accuracy with interpretability options:** The primary model provides highly accurate predictions by capturing complex patterns in applicant data, enabling more precise Credit Risk assessment. In parallel, a secondary, more interpretable model offers clear insights into the factors driving risk, supporting transparency and explainable decision-making when needed.

Overall, the deployment of this prediction model will strengthen our lending decisions, reduce operational costs, and support sustainable portfolio growth. I would be happy to discuss how this can be integrated into our existing workflows and the potential next steps for implementation.

Warm regards,
Chee Siong