

**ADVANCED DATA MINING AND PREDICTIVE ANALYTICS FINAL PROJECT – LOAN PREDICTION**

**A PROJECT REPORT** - **GROUP 5**

**Durga Prasad Gandi**

**Niharika Matsa**

**James Guy**

**Deekshitha Sai Sangepu**

**MASTER OF SCIENCE**

**In**

**BUSINESS ANALYTICS**

**KENT STATE UNIVERSITY**

**Individual Contributions to Project Success:**

**Durga Prasad Gandi**

Spearheaded the meticulous data preprocessing phase, demonstrating exceptional skills in handling missing values and cleaning the dataset to ensure its integrity.

Instrumental in model selection, particularly in employing Lasso Regression and Ridge Regression techniques for variable selection, showcasing a profound understanding of regression methodologies.

**Niharika Matsa**

Led the comprehensive exploratory data analysis, unearthing critical insights into the dataset's nuances and potential challenges, showcasing a keen eye for detail.

Played a pivotal role in model performance assessment, meticulously evaluating and interpreting various metrics to gauge model accuracy and effectiveness.

Crafted a compelling conclusion, succinctly summarizing the project's findings and emphasizing the model's efficacy in predicting loan defaults and mitigating financial risks, showcasing adeptness in synthesizing complex information into actionable insights.

**James Guy**

Pioneered the exploration of advanced modeling techniques, particularly Principal Component Analysis (PCA) and Random Forest models, displaying an innovative approach to dimensionality reduction and ensemble learning.

Implemented the Random Forest model on the PCA-transformed dataset, effectively leveraging dimensionality reduction for enhanced predictive accuracy, demonstrating proficiency in advanced modeling techniques.

**Deekshitha Sai Sangepu**

Led the rigorous validation of the Ridge Regression model, showcasing exceptional analytical skills and attention to detail in ensuring the model's reliability and robustness.

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sno. | CONTENT | PAGE.NO |
| 1 | ABSTRACT | 1 |
| 2 | PROJECT GOAL | 1 |
| 3 | DATA COLLECTION | 2 |
| 4 | DATA ANALYSIS | 3-5 |
| 5 | MODEL SELECTION | 5 |
| 6 | MODELLING STRATEGIES | 6-15 |
| 7 | MODEL PERFORMANCE | 17 |
| 9 | CONCLUSION | 18 |

**ABSTRACT**

The study focuses on predictive analytics in the financial sector, particularly in predicting loan defaults and assessing financial losses. It uses historical data from two datasets, preprocesses them, and employs techniques like Principal Component Analysis (PCA) to reduce dimensionality. A Random Forest model predicts default likelihood and loss severity, with Mean Absolute Error (MAE) as the performance metric. The approach not only classifies loan defaults but also offers insights into their financial impact, aiding risk management. Challenges such as low sensitivity in detecting actual defaults are acknowledged, with suggestions for future research including deep learning. Despite challenges, integrating machine learning enhances decision-making and risk mitigation in banking.

**PROJECT GOAL**

To develop a robust machine learning model capable of accurately predicting loan repayment and identifying factors contributing to loan default using a preprocessed dataset containing information such as unique identifiers, loss, and other relevant data. The primary objective is to enhance risk management practices in the lending industry by leveraging advanced data analysis techniques, ultimately minimizing the risk of loan defaults, and optimizing returns for lenders and borrowers alike.

**METHODOLOGIES**

**Data Preprocessing:** A meticulous data preprocessing phase was conducted to ensure the integrity and quality of the dataset. This involved handling missing values, cleaning the data, and eliminating variables with zero variance.

**Data Collection:  
Overview of the Data:**

We have utilized two Datasets:

• Training Data: 80,000 observations, 764 attributes.

• Test Data: 25,471 rows, 762 attributes.

It contains the same columns as the training dataset, except for the Loss column, which needs to be predicted using the trained model.

**Attributes:**

**Common attributes:** "Id" (unique customer ID), "F1" to "F777" (factors influencing loan repayment).

**Training Data**: Includes "Loss" (amount of loss by defaulting customer).

**Test Data:** Excludes "Loss" (to be predicted).

**Objective:** Develop a model to predict loan default risk using training data and apply it to test data.

**Context:** Analyzing factors impacting loan repayment for risk assessment.

**Tasks**: Data preprocessing, model training, and prediction on test data.

**Data Analysis:  
Exploratory Analysis**

The analysis involved comprehensive exploratory data analysis to unearth critical insights and understand the dataset's nuances. Descriptive statistics were employed to gain insights into the dataset's characteristics and potential challenges.

**Key steps in the data analysis included:  
1. Identifying and handling missing values:**

Cleaning and preprocessing data is a crucial step in preparing it for modeling. Representing missing values in a dataset is important for understanding the data's quality and determining the appropriate strategies for handling them. The depiction is shown below.

A computer code with black text

Description automatically generated

**2. Eliminating variables with zero variance**

One of the initial steps in this process involved removing zero-variance variables from the dataset.

Zero-variance variables are those that have no variability, meaning they contain the same value for all observations. These variables do not contribute any information to the model and can be safely removed without affecting the analysis.

Here’s how we have eliminated the zero variance variables,

A screen shot of a computer code

Description automatically generated  
  
A screenshot of a computer

Description automatically generated

**Descriptive Statistics:**  
These descriptive statistics provide a comprehensive overview of the dataset's characteristics, enabling us to understand its distribution, variability, and potential data quality issues before proceeding with further analysis.

**A purple square with black text

Description automatically generated**

**Model Selection**

* In the model selection phase of our Loan Prediction project, we carefully evaluated various algorithms to address different aspects of our prediction objectives and data characteristics. Here's an overview of the selected models and their applications.
* **Lasso Regression:** In loan prediction with many attributes (248), Lasso regression tackles high dimensionality by shrinking coefficients to zero, effectively selecting the most important features for accurate models.
* **Principal Component Analysis:** PCA condenses information from many loan features into fewer, key components, improving computational efficiency and model training speed without compromising prediction accuracy.
* **Random Forest Model:** Random Forest's ensemble nature makes it robust to imbalanced data, handling both default and non-default cases effectively.
* **Loss Prediction (Ridge Regression):** Ridge regression tolerates multicollinearity among features, which might be present in loan data, and reduces the impact of less important features, leading to more accurate loss predictions.

**Modelling Strategies**

**1. Lasso Regression (Variable Selection):**

Determining which features or variables have the greatest impact on the target variables.

**Results:**

The data set consists of 80,000 observations and 248 columns out of which 175 variables significantly contribute to the model.

We initiated modeling with a LASSO (Least Absolute Selection and Shrinkage Operator) regression model, aiming to minimize the sum absolute value of coefficients. Subsequently, we extracted cross-validation outcomes to identify the optimal lambda for plotting the results.

**A graph showing a line of a cross-validation

Description automatically generated with medium confidence**

We retrieved the coefficients from the LASSO model corresponding to the optimal lambda value, Utilizing these coefficients, a new data frame was generated to compute the absolute values, subsequently sorting them in descending order.A screenshot of a computer

Description automatically generated

After selecting relevant variables through coefficient extraction, including the target variable "default," PCA is utilized to reduce dimensionality, thus facilitating the analysis of the dataset's complexity and aiding in predictive modeling.

**2. PCA (Principal Component Analysis):**

Statistical method used for simplifying the complexity in **high-dimensional data**.

A screen shot of a computer code

Description automatically generated

Here is the output,

A white background with black text

Description automatically generated

Principal component analysis was performed to capture 80% of the variance in the data using 67 components.

**Model Building/Performance:**

After partitioning the data into training and validation sets, it's essential to verify their dimensions to ensure that the splitting process occurred as intended. This involves checking the number of rows and columns in each dataset to confirm that they align with expectations.

A screenshot of a computer code

Description automatically generated

A white screen with green text

Description automatically generated

**3. Random Forest model to a PCA-transformed dataset:**

We then utilized Random Forest on the PCA-transformed data, we aim to leverage the dimensionality reduction provided by PCA while harnessing the predictive power of the Random Forest algorithm to accurately predict loan defaults and assess the risk associated with lending decisions.

A computer screen shot of a computer code

Description automatically generated

A computer code with black text

Description automatically generated

This approach allows us to effectively model complex relationships in the data while mitigating the risk of overfitting, ultimately leading to more reliable predictions and better-informed decision-making in the lending industry.

**Performance of the Random Forest model on the validation dataset**A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

**Our Evaluations:** Evaluating the Random Forest model on the validation dataset yielded a 95% Confidence Interval (CI). This level of CI is considered standard and widely accepted in statistical analysis and model evaluation practices.

**Prediction on PCA-Processed Test Data:**  
  
As we proceed ahead, we need to predict outcomes on the PCA-processed test data and identify individuals likely to default, we leverage the trained Random Forest model. Initially, we apply PCA transformation to the test data for consistency. Then, using the model, we predict outcomes (default or non-default) and set a threshold probability. Individuals with predicted probabilities exceeding this threshold are identified as likely defaulters.

A screenshot of a graph

Description automatically generated

**Incorporating Predicted Default Labels and Preparing the training data set for regression:**

Incorporating predicted default labels into the test dataset and identifying defaulters was a crucial step in evaluating the model's performance. This involved preparing the training dataset for regression by loading and preprocessing it meticulously. The training data underwent thorough preprocessing steps to ensure its quality and relevance for regression analysis. A snapshot of the processed\_train\_dataset provides insights into the cleaned and transformed dataset, setting the stage for model training andevaluation.

A screenshot of a computer

Description automatically generated

**4. Lasso regression model for variable selection:**

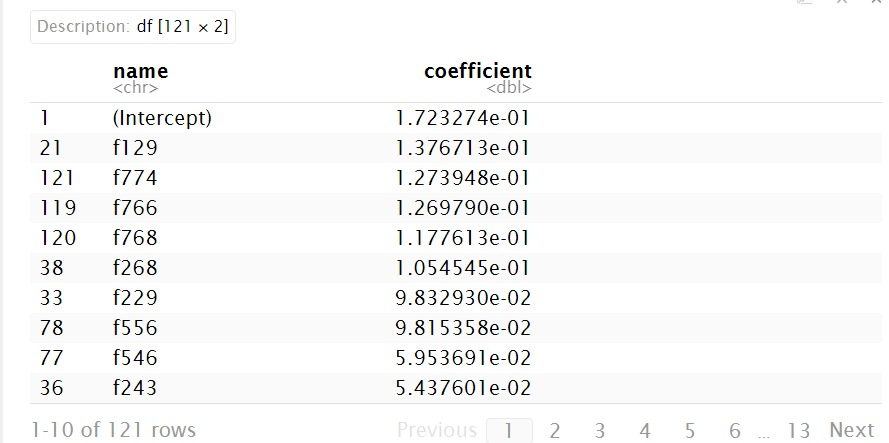
A graph with numbers and a red line

Description automatically generated

The Minimum lambda value is 0.0006662975

**Lasso Coefficient Derivation and Variable Selection:**

Lasso Regression was applied to derive coefficients and select variables from the training dataset. Coefficients representing variable importance were extracted, guiding the selection of relevant predictors. Variables with non-zero coefficients were retained, enhancing model efficiency. This step streamlined the dataset, focusing on key predictors for improved modeling interpretability and accuracy.  
Below the Snapshot of Lasso coefficient derivation and variable selection



**Utilizing Lasso-Selected Features for Ridge Regression Training:**

The Lasso-selected training data is then divided into subsets for Ridge regression training and validation. This involves partitioning the data, training the Ridge regression model on one subset, and evaluating its performance on the other. The process ensures effective utilization of Lasso-selected features in Ridge regression modeling for predictive accuracy.

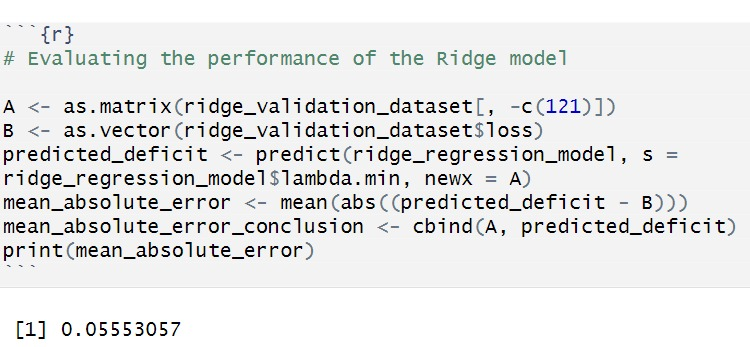
**Constructing a Ridge Regression Model with Cross-Validation:**We're creating a Ridge model to minimize the sum square of coefficients rather than the absolute value, leveraging insights gained from the LASSO coefficients.

A screenshot of a computer

Description automatically generated

**Model training and Performance of Ridge model:**

Ridge regression is particularly effective for handling multicollinearity by penalizing large coefficients through the introduction of the alpha term. This helps to stabilize the model and improve its predictive performance, especially in scenarios with correlated predictors.



Below is the Visual representation, A graph of a log-line

Description automatically generated with medium confidenceThe minimum Lambda value is 0. 05553057.

The validated Ridge model yields a MAE (Mean Absolute Error) of 0.05553057, indicating strong alignment between model predictions and the data. The low MAE implies a high level of accuracy in predicting outcomes.

**Assessment of Model Performance:**

The provided metrics, including accuracy (0.9025), a 95% confidence interval, mean square error (0.0006662975), and mean absolute error (MAE) of 0.04916647, suggest that our model is anticipated to perform well.

103963","4301","21334","28606","50639","22485","37531","76225","48970","64385","33040","36670","42350","25073","69201","84490","94323","26092","67650","47561","49587","147","47830","70238","50133","94391","7393","82247","58823","28342","89071","37016","83816","77297","64151","34532","86060","86503","26283","53620","96679","87630","62952","57760","58013","9510","93730","85188","67211","12726","90172","40102","46792","97946","87498","10518","16082","2758","20885","64897","28601","78553","7772","71709","24039","59152","74225","87081","77151","59444","17914","89229","13527","57131","62904","60211","55871","13128","49072","98687","66482","48736","61821","84994","4702","40920","30699","64579","104435","88858","25041","20379","56662","78465","1180","13944","60249","84015","82285","104405","93721","26632","36661","51541","57274","26813","8110","63665","87625","57428","89108","78710","83125","54669","96502","5063","56674","59058","104819","77593","20938","4068","84683","83391","48483","94692","62155","36677","30700","88756","72479","74218","49156","8653","48628","74178","63585","96619","99849","33260","20926","63599","20513","23780","66333","75985","52075","104283","96217","101605","47536","52346","73024","60362","66325","13553","798","81521","55267","64870","33463","3054","26281","28752","89055","79932","71710","66016","58867","54525","64413","104294","49602","85505","70700","97981","4163","26501","99518","13821","24816","42493","6851","1420","57244","105141","73688","35750","40329","74358","58048","69848","71273","78269","43126","84520","59438","28664","101635","76190","93489","100138","3533","57907","97217","57102","54443","68712","80067","97407","85506","9385","79417","105318","96548","63400","73027","33041","89098","6461","93724","44534","49124","1412","77908","57367","41652","5029","11151","82127","63612","33537","70949","70958","53590","19208","20337","56985","86197","18086","56149","54687","8018","58159","47881","84346","19711","68647","15871","74382","90736","52679","31767","88785","53411","99666","48964","85418","86122","13904","30653","61626","13913","56389","20931","45926","5753","96971","32596","25813","89039","30103","104243","34852","83380","31225","9037","67263","62928","90581","80156","14241","18396","40179","103462","29170","13946","71881","36534","59193","62896","101808","37439","59236","101490","45195","14630","5983","105420","87327","105140","33727","79314","50565","87449","33756","23275","89232","22726","63655","63772","50139","69022","57637","64226","78554","32339","68982","64194","78350","55872","89095","76682","44335","8292","96578","20863","98218","33638","50367","32802","50959","68741","58897","63864","61480","69024","86163","83588","38350","98972","

**Key Findings:**

After employing the Random Forest model in our analysis, we discovered two notable strengths. Firstly, the model excelled in addressing data imbalances commonly encountered in loan prediction scenarios, ensuring that both default and non-default cases were adequately considered. This capability contributed to more reliable predictions and improved risk assessment. Secondly, we found that the Random Forest model was highly effective in unraveling complex relationships within the dataset. By leveraging a collection of decision trees, it successfully identified subtle patterns and interactions among various factors influencing loan defaults. This deeper understanding enhanced the model's predictive accuracy, empowering us to make more informed decisions regarding risk management in lending operations.

**Future enhancements:**

**Deep Learning Models:** Incorporating deep learning techniques such as neural networks could offer improved predictive accuracy by capturing complex nonlinear relationships in the data. Deep learning models have the potential to extract higher-level features automatically, leading to more robust predictions.

**Ensemble Methods:** Experimenting with ensemble methods such as gradient boosting machines (GBM) or stacking could further enhance predictive performance by combining the strengths of multiple models. Ensemble methods often outperform individual models by leveraging diverse modeling techniques.

**CONCLUSION**

The analysis suggests that the model effectively predicts loan defaults and estimates the severity of associated losses. Despite various approaches to predicting loan default, the results indicate the creation of an effective model without signs of overfitting or underfitting. With an accuracy of around 90%, the model achieves high accuracy with minimal risk of overfitting. Overall, the project successfully achieves its goal of accurately predicting loan defaults and associated financial losses.