Course Code: BC3409

Assignment: Homework 2

Submitted by: Shane Lim

# Results for Logistic Regression, Decision Tree, Gradient Boosting and Neural Network

**Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
| **Logistic Regression** | **Decision Tree** | **Gradient Boosting** | **Neural Network** |
| **0.972** | **0.957** | **0.956** | **0.9665** |

**Confusion Matrix**

A screenshot of a graph

Description automatically generated

## Discussion on results:

1. Looking at the accuracy metrics only, logistic regression provides the most accurate results (0.972) and Gradient Boosting offers the least accuracy (0.956). Overall assume we have a minimal accuracy requirement of 95%, all of the models we have used will be effective in identifying default rates
2. However, Logistic Regression performs very well for classifying non-defaults (0s), it struggles with defaults (1s), where it misses 69%~ of the defaults (46 out of 67 defaults). This inaccuracy could be attributed to the **imbalanced dataset** with fewer default cases
3. Ultimately, if we want to prioritise the models which can detect defaults we should go with Decision Tree or Gradient Boosting

# Loan Default Answers

## Explain the pros and cons about your model, including limitation (can be both quantitative and qualitative).

**Logistic Regression:**

* **Pros**:
  + **Easy to Understand and Interpret**: Logistic regression models are easy to explain and interpret. This is useful when explaining to non-technical stakeholders. For example, the coefficient for each of the independent variables explains how it affects the likelihood of loan default.
  + **Efficient**: Logistic regression requires minimum amount of computational power to train.
  + **Good for Linearly Separable Data**: Performs well when there is a linear relationship between the features and the target variable.
* **Cons**:
  + **Linear Assumptions**: One key assumption of Logistic regression is that the independent variables and log odss of the dependent variables (in this case the probability of a loan default) have a linear relationship. However, in most real world applications these relationships or non-linear.
    - Logistic Regression will underperform when there are complex relationships between variables
  + Reduction in performance when independent variables have average or above average multicollinearity.

**Decision Tree:**

* **Pros**:
  + **Non-Linear Relationships**: Decision trees are excellent at capturing non-linear relationships between independent and dependent variables because the model splits the data into subsets based on certain conditions. (Unlike Logistic Regression which makes assumption of linearity)
  + **Highly Interpretability**: For non-technical stakeholders, it is easy to understand the decision-making process of the model.
  + **No Need for Feature Scaling**: Decision trees do not require feature scaling because each data points are split by utilising a threshold. These thresholds are selected based on the relative ordering of the feature values.
* **Cons**:
  + **Overfitting**: Decision trees are very prone to overfitting, especially with small datasets. This will affect the reliability of the model is applied to unseen data.
  + **Instability**: Minor changes in the dataset can lead to completely different splits and trees.

**Gradient Boosting:**

* **Pros**:
  + **High Accuracy**: Because Gradient boosting involves sequentially creating one tree at a time and for every subsequent tree that is created more weight is placed on the predictions with errors. This leads to high predictive accuracy.
  + **Handles Non-Linear Relationships**: Similar to decision trees, gradient boosting can handle complex non-linear relationships well.
  + **Reduces Overfitting**: Due to ensemble learning and regularization the chances of the model overfitting are significantly lower.
* **Cons**:
  + **Computationally Intensive**: Substantial computational resources are required to train large datasets.
  + **Less Interpretable**: Although Gradient Boosting more accurate, it requires some understanding of Statistics to understand how the model makes decisions.

**Neural Network:**

* **Pros**:
  + **Powerful Non-Linear Modeling**: Similar to Gradient Boosting Neural networks can identify highly complex and non-linear patterns in data. This is especially true for non-text data like images.
  + **Generalization**: Using sufficient data and appropriate regularization, we can manage the complexity of neural networks and thus generalize well to unseen data.
* **Cons**:
  + **Requires Large volumes of Data**: Neural networks typically need large amounts of data to perform well.
  + **Computationally Expensive**: When developing Neural networks models it is financially costly and time-consuming compared to traditional models.
  + **Lack of Interpretability**: Neural networks are often considered "black boxes".

## How to overcome the weakness of your model (future study).

**Logistic Regression**:.

* Consider using variations to the basic logistic regression model like polynomial logistic regression or regularization (Lasso or Ridge) to improve the model's performance.

**Decision Tree**:

* Apply pruning to reduce branches that do not contribute to predictive power and thus prevent overfitting of the training data
* When the dataset is small apply cross-validation to reduce overfitting.
* Explore ensemble methods like Random Forest or Gradient Boosting to stabilize and improve accuracy.

**Gradient Boosting**:

* Reduce computational resources required by using distributed learning techniques or parallelization.
* Apply hyperparameter tuning to optimize model performance.

**Neural Network**:

* Add more data or introduce more layers or nodes to improve model performance.
* Consider using dropout or batch normalization which refers to normalization of inputs to have a mean activation output of zero and unit standard deviation. These methods regularize the model and prevent overfitting.

## Any descriptive analysis you could think of for this case. Example, confidence interval.

* **Summary Statistics:** 
  + Summary statistics provides an overview (mean, s.d. etc.)
* **Box and whiskers plot:** 
  + Used to identify the distribution of the data points  through their quartiles
* **Confidence Interval**:
  + Users can calculate the confidence intervals for the accuracy of each model to understand the variability of performance across different data samples.
* **Correlation Analysis**:
  + Conduct correlation analysis on the bank balance, annual salary, and employment status to measure how strongly the independent variables are associated with loan defaults.
* **Class Distribution**:
  + I would use a histogram to identify and handle the class imbalance. The datasets contain much more non-defaults compared to defaults.

## What historical data variables are considered most influential in predicting loan defaults, and how are they weighted in your analysis?

A black screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

Based on the **feature importance** from Logistic Regression, Decision Tree and Gradient boosting:

* **Bank Balance** is the most influential feature in predicting loan defaults for Decision Tree amd Gradient Boosting:
  + **Decision Tree**: Importance score of **0.676578**, meaning it has the most influence on predictions.
  + **Gradient Boosting**: Bank balance ranks the highest with a score of **0.907444**.
* **Annual Salary** is also significant but less influential compared to bank balance:
  + **Decision Tree**: Importance score of **0.304266**.
  + **Gradient Boosting**: Importance score of **0.088757**.
  + **Employment Status** (whether employed or not) is less influential but still matters. **Logistic Regression** suggests that Employment Status is a more influential factor. A coefficient of **-3.092** suggests that if a person **is employed**, the **log-odds of default** **decrease by 3.09**

## How do economic indicators and market trends impact the accuracy of your loan default predictions, and what strategies are in place to adapt to changing conditions?

* **Unemployment rates**:
  + A rising unemployment rate could lead to higher loans because they are do not have a steady source of income to repay their loans
* **Interest rates**:
  + Higher interest rates increase loan costs, which may increase default rates for individuals with high number of variable-rate loans.
* **Inflation**:
  + When living costs increase it reduces the loanee’s disposable income which could lead to more defaults.

**Adaptation Strategy**:

* **Incorporate Macroeconomic Variables**: We can improve model performance by including external macroeconomic indicators like unemployment rates or inflation as independent variables in our model.
* **Dynamic Modeling**: Utilize time-series models or adaptive machine learning techniques to update predictions based on the market conditions.

## How do you balance the need to mitigate default risk with the goal of providing access to credit for underserved or high-risk borrowers? Are there any ethical considerations in this decision-making process?

* **Mitigating Risk:** 
  + Instead of outright rejecting loans to high-risk borrowers, banks could adjust loan terms (e.g., higher interest rates, shorter loan periods). For example, Baidu was interested in offering small loan to retail customers buying products from their platform. Unlike most developed countries, the risk with lending in the Chinese market is that less than 20% of people have credit profiles or credit ratings. Thus to accurately assess the borrowers’ reliability, Baidu worked together ZestFinance to make the decision based on factors like purchase history.
* **Ethical Considerations:**
  + Fairness: The model should avoid discriminating against certain demographic groups. Ensure fairness by regularly auditing the model for bias. Additionally, from my internship experience in a bank, there is always still a human touchpoint even though the entire process can be done by AI/ML.
  + Transparency: Data Scientist should be able to explain the algorithm behind the model’s decisions to borrowers so they understand why they were approved or denied.

## The prescriptive analysis on loan default aims to enhance decision-making, reduce default risk, and optimize lending practices while maintaining a balance between profitability and risk mitigation. How do you think accurately predicting loan default can help in any decision making? (The importance of your model to the bank.)

* **Risk Management**: Accurate prediction of loan defaults helps the bank minimize losses by identifying high-risk borrowers using variables such as debt-to-income ratios, credit delinquencies, property values and adjusting loan terms interest rates, and collateral requirements, minimizing their exposure to defaults (e.g., lower interest rates, higher credit limits). By accurately predicting defaults, banks can refine their lending practices to target niche segements of borrowers who might have been previously difficult to lend to, potentially improving profitability while minimizing default risk.
* **Resource Allocation**: The bank can allocate more resources to collections or recovery efforts for high-risk loans, while focusing marketing efforts on low-risk customers. Furthermore, it will reduce the need for the manual processes involved in loan approval and underwriting, such as staffing, training, and compliance. This can improve the overall profitability of the bank by freeing up resources for more impactful work.
* **Regulatory Compliance**: Accurate default prediction ensures the lender maintains healthy risk-weighted assets, complying with regulatory standards like Basel III. This will build trustwith stakeholders and regulators and reduce future risks from legal and reputation issues