DUKE FUQUA SCHOOL OF BUSINESS

Data Science Final Team Project

Credit Risk Calculation & Approval Scheme

Team 3

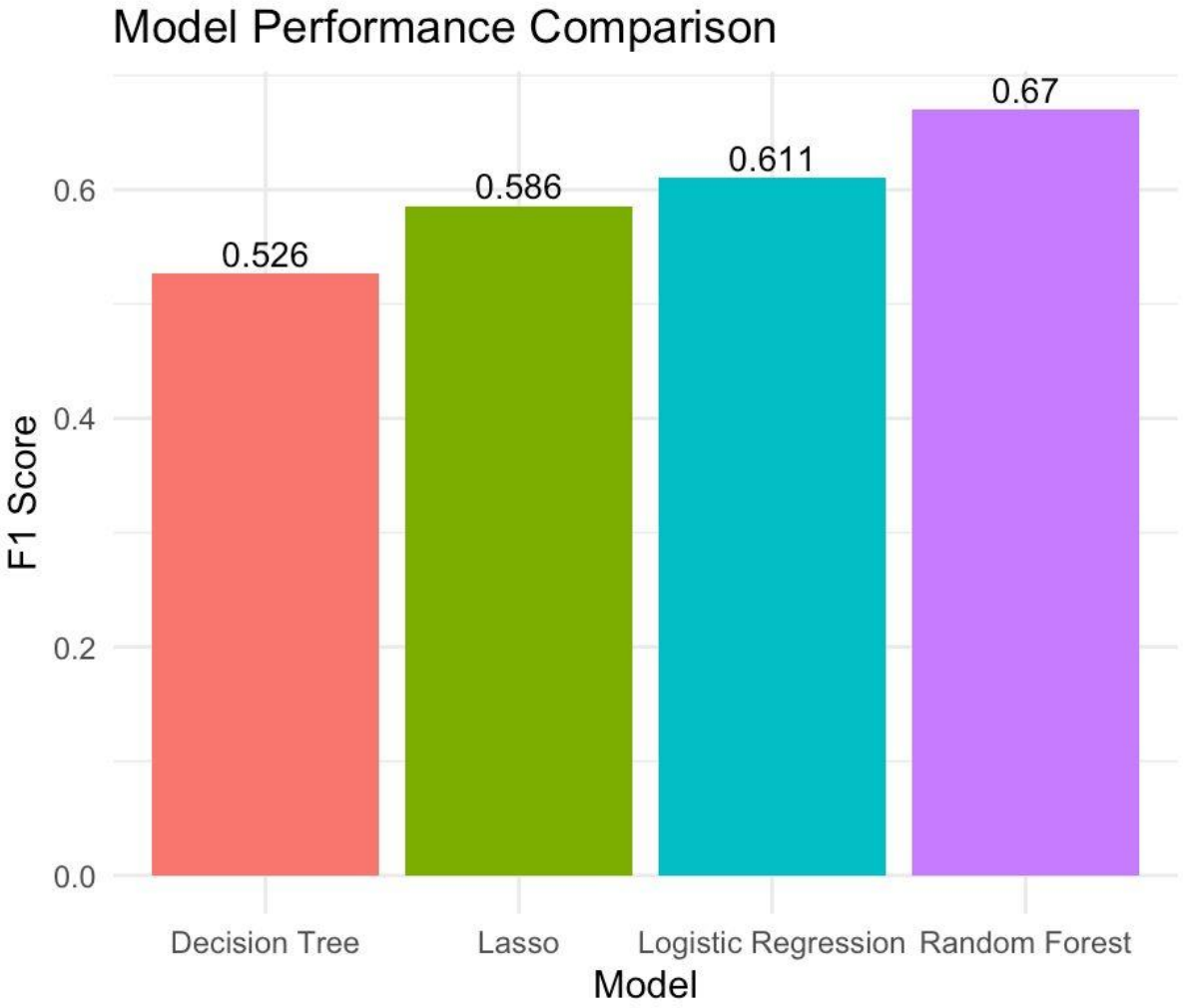
Mia Kwon, Zhixuan Li, Jason Liao, Yitian Wang, Sihan Zhou

Executive Summary

DUKE FUQUA SCHOOL OF BUSINESS

- **Aim**: design a predictive model to assist making informed decisions about credit approvals. Extend credit facilities to deserving applicants and safeguarding issuer's interests
- **Chosen Models:** Five Logistic Models(max F1=0.61), LASSO Model (F1=0.58), Classification Tree(F1=0.68), Random Forest(F1=0.72), K-means Clustering. All conducted with cross validation to prohibit over-fitting.
- **Conclusion**: Classification models tend to do better in binary prediction. Random Forest demonstrated the best performance in terms of accuracy and balance between precision and recall, as indicated by its highest F1 score among the models we tested.

*F1 Score a metric used to evaluate the performance of binary classification models, especially when the classes are imbalanced.



Model Performance Comparison

Business Understanding

DUKE FUQUA SCHOOL OF BUSINESS

- Credit card companies have always been a key player in the economy by helping manage the risks of missed or late payments. Due to the importance of credit system, making wise credit approvals is critical for banks.
- In this case, our aim is to design a predictive model to assist credit card companies in making informed decisions about credit approvals. Our model compiled detailed profiles based on an individual's personal traits, credit history, and other relevant features.
- We want to identify candidates who are likely to maintain a healthy credit relationship, thereby seeking a balance between extending credit facilities to deserving applicants and safeguarding the issuer's interests.

Data Preparation

- Create A Responsive Variable

Binary(1:”Not Approved”, 0:”Approved)

Tracked customers’ payment status records since account open date, and chose customers who overdue for more than 60 days as target customers.

- Clean Feature Variables

Binary variables were assigned numeric values and continuous variables such as age and others were adjusted. Lastly, categorical variables were transformed into numeric or dummy variables.

- Handle Imbalance

We fixed dataset imbalance by expanding the "not approved" dataset and using models sensitive to minorities. The goal was to aim for accurate predictions beyond the majority class.

Modeling - Logistic and LASSO

DUKE FUQUA SCHOOL OF BUSINESS

Logistic Model

- 5 different types of logistic models based on
  - Financial Stability
  - Family Status
  - Loan and Credit History
  - Demographic Information
  - Asset Ownership

- Relative sensitive to minority, good for a unbalanced sample

- Avoid overfitting by 5 folds cross validation

Performance Evaluation

- Loan and Credit History achieved highest F1-Score (.60)

LASSO Model

- Goal: Avoid over-fitting
- Find optimal Lamda with cross validation to balance model complexity and training data fit
- Tried to omit highly intercorrelated variables that are less related to Approval status, not effective

Performance Evaluation

- F1-Score: 0.58
- Implication: better than random guessing, but away from excellence.
- LASSO might not be the best model for binary prediction

Modeling - Classification Models

DUKE FUQUA SCHOOL OF BUSINESS

Decision Tree

- Balanced the factor class imbalance of target variable using ROSE method
- Reduced overfitting by 10 folds cross validation

Performance Evaluation

The best F1 score for Decision Tree model was 0.68.

Random Forest

- Aggregates multiple decision trees
- Chose to enhance predictive accuracy and reduce overfitting
- Reduced overfitting by 10 folds cross validation

Performance Evaluation

- The best F1 score for Random Forest model was 0.72.
- Random Forest model turned out to be the best fitted model for our prediction.

Deployment

Implement the optimized model (Random Forest) in the credit card process for instant decisions. Regularly update and monitor the model for consistent performance as we consider the changing data patterns.

Ensure correct use of the model to avoid biases against groups based on attributes like race or gender, to prevent legal or reputational risks.

Constant monitoring is important for efficient modeling

Key Risks:

- Model Drift: Regular updates for consistent performance
- Over-reliance: Have human oversight to avoid systematic errors
- Data Privacy: Adhere to data protection laws and appropriate practices
- Model Bias: Conduct fairness checks and make necessary adjustments