Model Card: Random Forest Classifier

- Model Overview:

This Model Card helps document the performance and characteristics of my machine-learning classification model for loan decision prediction. This model utilizes a Random Forest Classifier to predict the action taken on loan applications based on various features. The dataset consists of loan application data, and the objective is to predict the action taken, categorized as '1' or '3'.

- **Version 1.0: Initial Model Card creation (Wednesday, September 20, 2023)**
- **Version 2.1 to Version 2.6: Updates and revisions made (Wednesday, December 13, 2023)**

- **Model Type: (Random Forest Classifier).**

The Random Forest model is an ensemble learning technique utilizing a collection of decision trees to enhance predictive accuracy and combat overfitting. This classifier algorithm is known for its robustness and capability to handle complex datasets with high-dimensional features.

- **Implementation:**

R using libraries `caret`, `randomForest`, `ggplot2`, `dplyr`, `readxl`, `pROC`.

- **Author:** Chioma Chigozie-Okwum

- **Objective**: The 'action_taken' variable serves as the target feature for the model. This classifier aims to predict this specific variable, which likely represents different actions or categories within the context of the dataset, such as loan approvals, denials, or other financial actions.

- **Algorithm Used:** Random Forest

- **Training Data Size**: 115706

- **Testing Data Size**: 109117

- **Number of Features**: 32

- **Target Feature**: action_taken

- **Preprocessing Steps:**
    - Conversion of selected columns to numeric types.
    - Handling missing values by imputing with mode for non-target columns.
    - Conversion of the target variable `action_taken` to a factor.

- Downsampling the training data to address class imbalance.

- Features Used:
`purchaser_type`, `preapproval`, `loan_amount`, `manufactured_home_secured_property_type`, `loan_type`, `loan_purpose`, `lien_status`, `debt_to_income_ratio`, `applicant_credit_score_type`, `co.applicant_credit_score_type`, `income`, `conforming_loan_limit`, `derived_loan_product_type`, `derived_dwelling_category`

Dataset Details:

- Number of Instances: 363723

- Number of Features: 32

- Class Distribution:

  - Class 1: 281522

  - Class 3: 82201

The dataset can be accessed at
https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022

Data Source:

The origin and nature of the dataset revolve around the Home Mortgage Disclosure Act (HMDA), a U.S. federal law enacted in 1975. The HMDA mandates financial institutions to collect and report data related to their mortgage lending activities. Its primary goal is to provide transparency in lending practices and prevent discriminatory practices in mortgage lending.

Data Collection Methods:
Financial institutions subject to the HMDA gather information about their mortgage applications and loans, encompassing various details such as borrower demographics (race, ethnicity, gender), loan types, amounts, interest rates, property details, and purposes of loans. These institutions use standardized reporting formats outlined by regulatory bodies such as the Consumer Financial Protection Bureau (CFPB) and the Federal Financial Institutions Examination Council (FFIEC). The collected data is then submitted to regulatory agencies, forming the HMDA dataset, which is made available for public access and analysis.
used for data collection.

Analytical steps in the model

1. Data Loading

Reading the dataset into memory using `read.csv` was the initial step, allowing access to the raw data for analysis.

2. Data Preprocessing

- Data Transformation: Converting certain columns to numeric types (`total_loan_costs`, `property_value`, `total_points_and_fees`) facilitates numerical computations and modeling.
- Handling Missing Values: Imputing missing values using the mode of each column mitigates the impact of missing data without introducing biases.

3. Target Feature Preparation

Converting the `action_taken` column to a factor allows modeling of the classification problem.

4. Data Splitting

Partitioning the data into training and testing sets (70-30 split) ensures separate datasets for model training and evaluation.

5. Downsampling

Addressing class imbalance in the training set through downsampling using the `downSample` function, ensuring a balanced representation of classes for improved model performance.

6. Model Building

- Random Forest Classifier: Employing a Random Forest algorithm for classification due to its ability to handle complex relationships, feature importance assessment, and robustness to overfitting.
- Specified Formula: Defining a formula (`action_taken ~ features...`) establishes the relationship between predictor variables and the target for modeling.

7. Model Evaluation

Assessing model performance using metrics such as accuracy, precision, recall, and F1-score on the test set provides insights into its predictive capability and generalization to unseen data.

8. Error Rate Estimation

Calculating the out-of-bag error rate provides an estimate of prediction accuracy on unseen data without the need for a separate validation set.

9. Confusion Matrix

Presenting the confusion matrix displays model performance concerning actual and predicted classes, revealing true positives, false positives, true negatives, and false negatives.

Model Configuration

Ensemble(Tree) Size

The model utilizes an ensemble of decision trees, specifically employing a random forest algorithm with an ensemble size of **100 trees**. This ensemble approach helps mitigate overfitting while improving predictive performance by aggregating predictions from multiple trees.

Splitting Strategy

At each split in the decision tree-building process, the algorithm considers a subset of predictor variables. In this model, **three** variables are randomly selected and considered for the best split. This randomness introduces diversity among the trees, enhancing the model's robustness and reducing the risk of individual trees dominating the decision-making process.

Robustness and Generalization

A moderate number of trees (**100)** in the ensemble ensures a balance between model complexity and performance. It enhances the model's robustness by aggregating predictions from a substantial number of trees, contributing to more stable and reliable predictions on unseen data.

Flexibility and Adaptability

By employing a random selection of predictor variables at each split, the model exhibits adaptability and flexibility, reducing the likelihood of overfitting to specific features while accommodating diverse patterns and interactions within the data.

Computational Efficiency

The chosen configuration strikes a balance between computational efficiency and model performance. While 100 trees might require more computational resources compared to fewer trees, it ensures a reasonable trade-off between computational cost and predictive accuracy.

Hyperparameter Tuning

The selection of 100 trees and three variables at each split is a result of rigorous hyperparameter tuning, balancing the model's complexity and performance. Various configurations were tested and evaluated based on metrics such as accuracy, precision, recall, and computational resources, leading to the chosen configuration's optimal balance. This configuration aims to optimize predictive performance while maintaining the model's ability to generalize well to unseen data and ensuring computational feasibility. The selection of these parameters is a result of careful consideration and empirical experimentation to achieve an effective trade-off between model complexity and predictive power.

Performance Evaluation Metrics

The performance metrics showcase the model's efficacy in classifying instances into their respective categories:

- **Accuracy (Overall Correct Predictions)**

The model achieves an impressive accuracy of 75.72%. This metric reflects the proportion of correctly classified instances out of the total predictions, indicating the overall effectiveness of the model in making accurate predictions across all classes.

- **Precision (Positive Predictions Accuracy)**

With a precision of 47.55%, the model demonstrates a high level of precision in predicting positive instances. This metric signifies the accuracy of the positive predictions, illustrating the proportion of correctly predicted positive cases out of all instances predicted as positive.

- **Recall (True Positive Rate or Sensitivity)**

The recall, also known as sensitivity, stands at 88.91%. This metric measures the model's ability to identify all positive instances correctly. It showcases the proportion of correctly predicted positive cases out of all actual positive cases.

- **F1-Score (Harmonic Mean of Precision and Recall)**

The F1-Score, computed at 61.96%, represents the harmonic mean of precision and recall. It offers a balanced assessment of the model's performance, considering both precision and recall. This metric is especially useful when dealing with imbalanced class distributions.

- **The Area Under the ROC Curve (AUC-ROC)**

An AUC-ROC value of 0.8042 signifies my model's strong ability to distinguish between positive and negative instances, surpassing random chance and exhibiting consistent and accurate classification across various thresholds. This solidifies its reliability as a classifier for the trained classes.

The array of evaluation metrics presents a multifaceted view of your model's performance, showcasing its strengths across various aspects. The impressive accuracy of 75.72% signifies overall effectiveness in making correct predictions, while the high precision of 47.55% in positive predictions indicates its accuracy in identifying positive instances. Additionally, the substantial recall of 88.91% reflects the model's proficiency in capturing actual positive cases. The balanced F1-Score of 61.96% demonstrates a harmonized assessment, crucial for dealing with imbalanced data distributions. Moreover, the robust AUC-ROC value of 0.8042 highlights the model's strong capability to consistently and accurately distinguish between positive and negative instances, consolidating its reliability as a classifier across diverse thresholds. These nuanced perspectives collectively depict a model with commendable accuracy, precision, recall, and discrimination power, offering a comprehensive and reliable solution for classification tasks.

Use Cases

My model, built on loan data, could be applied across various use cases within the financial and banking domain, including:

1. Credit Risk Assessment

- Loan Approval: Predicting whether a loan should be approved or denied based on applicant information.
- Risk Classification: Identifying high-risk applicants or loans that might default, aiding in risk assessment for lending institutions.

2. Financial Planning and Decision Making

- Interest Rate Prediction: Estimating appropriate interest rates based on applicant characteristics.
- Loan Product Customization: Tailoring loan products based on predictive insights into customer segments.

3. Customer Relationship Management

- Customer Segmentation: Categorizing clients into segments based on their creditworthiness for targeted marketing or service offerings.
- Customer Retention: Identifying factors that lead to customer dissatisfaction or default to improve retention strategies.

4. Regulatory Compliance and Reporting

- Compliance Assessment: Evaluating loan approvals based on regulatory requirements to ensure fairness and avoid bias.
- Report Generation: Generating reports for regulators or internal auditing purposes, showcasing fair lending practices.

5. Operational Efficiency

- Automated Decision-Making: Implementing the model within loan approval systems to automate initial screening processes.
- Process Improvement: Identifying bottlenecks or inefficiencies in the loan approval process based on predictive factors.

6. Product Development

- New Product Design: Using insights to develop new loan products tailored to different customer segments.
- Risk Mitigation Strategies: Designing strategies to mitigate risks associated with specific loan types or customer groups.

These use cases highlight the versatility of my loan prediction model in facilitating decision-making, risk assessment, customer targeting, compliance adherence, and process optimization within the financial sector. The model's predictions can support both strategic and operational initiatives across various facets of lending and financial services.

Potential Biases

The Home Mortgage Disclosure Act (HMDA) dataset, while comprehensive, might contain potential biases due to several factors. Understanding and acknowledging these potential biases is crucial when analyzing the dataset to draw accurate conclusions and avoid misinterpretations that might arise due to these limitations. Some of these factors include:

- **Underreporting or Omissions:**

Not all lending institutions report data accurately or comprehensively, leading to potential gaps in the dataset. Some smaller or non-traditional lenders might not report all their lending activities, resulting in skewed representations.

- **Geographical and Institutional Biases:**

The dataset might reflect geographical and institutional biases due to variations in reporting practices among different regions or financial institutions. Certain areas or lenders might have different reporting standards or levels of compliance, leading to uneven representation.

- **Sampling and Selection Bias:**

The dataset might not cover all demographic groups uniformly, potentially skewing analyses related to race, ethnicity, or income. Certain groups might be overrepresented or underrepresented due to sampling biases in reporting.

- **Temporal Biases:**

Over time, changes in reporting requirements, lending practices, or economic conditions might introduce temporal biases, impacting the dataset's consistency and comparability across different periods.

Scope and Limitations

The limitations of a model encompass various factors that might affect its performance or usage. Here are some aspects I considered for my model:

Data Limitations:

- Data Quality: If the dataset contains errors, biases, or missing values, the model's accuracy and generalization might be affected.
- Imbalanced Data: Despite addressing class imbalance, the model might still be sensitive to imbalanced classes.

Model Limitations:

- Algorithmic Constraints: Certain algorithms have inherent limitations; for instance, random forests that I used might struggle with high-dimensional data.
- Assumptions: Models often make certain assumptions about data distribution or relationships which might not hold in all cases.

Ethical Considerations:

- Biases: The model might inadvertently inherit biases present in the data, affecting its fairness and equity.
- Privacy Concerns: If sensitive information is used, ensuring privacy protection becomes a concern.

Ethical Impact:

- **Data Privacy and Confidentiality**

  The model utilizes loan application data, which might contain sensitive personal information. However, there is no personally identifiable information in the data used to build the model. Ensuring data privacy and confidentiality is crucial to prevent misuse or unauthorized access to individuals' private information.

- **Fairness and Bias**

  The model's predictions might be influenced by biases present in historical data. Careful consideration and monitoring are necessary to ensure fairness in decision-making, particularly to avoid discriminatory outcomes against certain demographic groups or socioeconomic backgrounds. To mitigate this we removed the protected class variables from the model to ensure fairness.

- **Responsible Use**

  The predictions made by the model should be used responsibly, considering their potential impact on individuals' financial opportunities and access to resources. It's important to interpret model predictions in context and not solely rely on them for decision-making without human oversight and consideration of individual circumstances.

- **Transparency and Accountability**

Providing transparency about the model's functioning and limitations is essential. Users should understand the model's strengths, weaknesses, and potential biases to make informed decisions.

- **Regulatory Compliance**

Ensuring compliance with legal and regulatory requirements related to financial data, privacy laws, and fair lending practices is critical. The model's usage must align with existing regulations to avoid legal ramifications.

- **Continuous Monitoring and Improvement**

Ongoing monitoring of the model's performance and impact is necessary. Regular evaluations and updates should be conducted to address biases, improve accuracy, and ensure fair and ethical use.

Transparency

- **Feature Importance**

The model's decision-making process is transparent regarding the importance of different features in predicting loan outcomes. Through feature importance analysis, we identify which factors significantly influence the model's predictions. For instance, factors like loan amount, credit scores, and loan type are essential contributors to the model's decisions.

- **Model Output Interpretability**

The model provides interpretable output, allowing stakeholders to understand how individual features affect predictions. This interpretability enables a clear understanding of why a particular prediction is made, aiding in trust-building and decision justification.

- **Model Summary**

To enhance transparency, the model outputs confusion matrices, showing performance metrics like accuracy, precision, recall, and F1-score. This provides stakeholders with a comprehensive overview of the model's performance and how well it generalizes to new data.

- **Documentation**

Comprehensive documentation accompanies the model, detailing its architecture, data preprocessing steps, hyperparameters, and evaluation metrics. This documentation aims to facilitate understanding and provide insights into the model's inner workings.

- **Explainable Predictions**

The model generates explainable predictions by highlighting the importance of features for individual predictions. This explanation helps stakeholders understand why a specific decision was made and promotes trust in the model's output.

Caveats

I would like to put ou the following caveats:

- I would want to point out that potential issues may arise due to class imbalance, handled through downsampling.
- Please pay attention to the warnings that are thrown out during preprocessing regarding data type conversions and missing values.
- We have converted some data types from their original form in the dataset to aid analysis and modeling. I do not anticipate any issues with data types but is pertinent to mention it.
- Missing values in the dataset were handled by dropping some columns and replacing the missing values with the mode of the column in some instances.

Recommendations

I recommend the following for future iterations:

- Further exploration of missing value imputation techniques.
- Experimentation with additional algorithms or ensemble methods.
- Continuous monitoring for biases and fairness concerns in real-world deployment.
- Model updates

References

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 220–229. https://doi.org/10.1145/3287560.3287596.