Case Study: Predicting Loan Default Risk Using Machine Learning

Overview

Managing loan default risk is a critical challenge for financial institutions, as it directly impacts profitability and credit stability. This project leveraged advanced machine learning techniques to predict loan default risk, identify key predictors, and provide actionable insights for improved lending decisions.

Objective

- Develop a predictive model to assess the likelihood of loan defaults.
- Uncover key factors driving default risk.
- Offer recommendations for financial institutions to enhance risk management strategies.

Tools and Techniques

- **Tools**: SQL, R, Tableau.
- **Techniques**: Logistic regression, Random Forest, Decision Trees, Exploratory Data Analysis (EDA), Feature Engineering.

Process

1. Data Collection and Preprocessing

- **Dataset**: Acquired from Kaggle, containing 1.5 million rows of application and historical loan data.
- **Initial Issues**:
  - 22% missing values.
  - Outliers in key financial variables like income and loan amount.
- **Actions Taken**:
  - Removed columns with more than 40% missing data.

    - Imputed missing values using median and mode methods.
    - Applied log transformations to address skewed variables.

2. Feature Engineering

- Aggregated and created new features to enhance model inputs:
    - **Address Mismatch Count**: Signaled borrower instability.
    - **Weighted Credit Bureau Inquiries**: Accounted for recency of inquiries.
    - **Social Circle Default Rate**: Captured social influence on borrower behavior.
- Generated metrics like Debt-to-Income Ratio and Loan-to-Value Ratio.

3. Modeling

- **Logistic Regression**:
    - Served as a baseline model.
    - Adjusted thresholds improved sensitivity to 69.92%.
- **Random Forest**:
    - Achieved superior performance with:
        - AUC = 0.9905.
        - Accuracy = 98.18%.
        - Sensitivity = 78.18%.
        - Specificity = 99.91%.
- **Decision Trees**:
    - Provided interpretable classification rules for practical decision-making.

4. Visualization

- Created interactive Tableau dashboards for insight presentation.
- Highlighted visualizations:
    - Feature importance analysis.
    - ROC curves showcasing model performance.
    - Default rate comparisons by demographic factors.

Results

- **Model Performance**:
  - Logistic Regression: AUC = 0.7392, Accuracy = 68.21%.
  - Random Forest: AUC = 0.9905, Accuracy = 98.18%.
- **Key Predictors**:
  - External credit scores (EXT_SOURCE_2, EXT_SOURCE_3).
  - Debt-to-Income Ratio.
  - Employment duration (DAYS_EMPLOYED).
- **Insights**:
  - Male borrowers exhibited higher default rates (10.68%) compared to females (7.84%).
  - Borrowers with high debt-to-income ratios and low external credit scores were at greater risk.

Recommendations

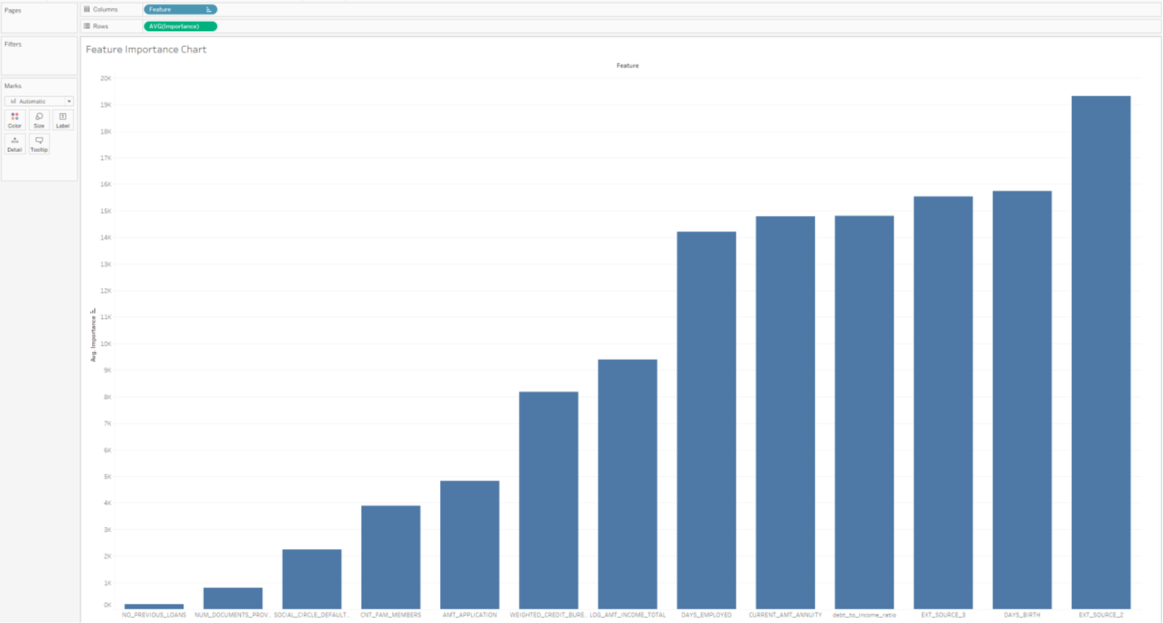
1. **Risk-Based Segmentation**:
   - Categorize borrowers into low, medium, and high-risk groups for targeted risk management.
2. **Automated Screening**:
   - Utilize predictive models to pre-screen loan applicants efficiently.
3. **Customized Loan Offers**:
   - Tailor loan terms and interest rates based on risk profiles.
4. **Model Refinement**:
   - Explore advanced techniques like Gradient Boosting to further improve model accuracy.
5. **Data Enrichment**:
   - Integrate additional datasets, such as behavioral and transactional data, for enhanced predictions.

Reflection

- **Challenges**:
  - Addressed class imbalance through weighted logistic regression and threshold adjustments.
  - Managed large datasets efficiently using SQL for preprocessing and integration.
- **Key Takeaways**:
  - Gained proficiency in advanced machine learning techniques and data visualization.
  - Demonstrated the ability to derive actionable insights from complex datasets.

Sample Visuals

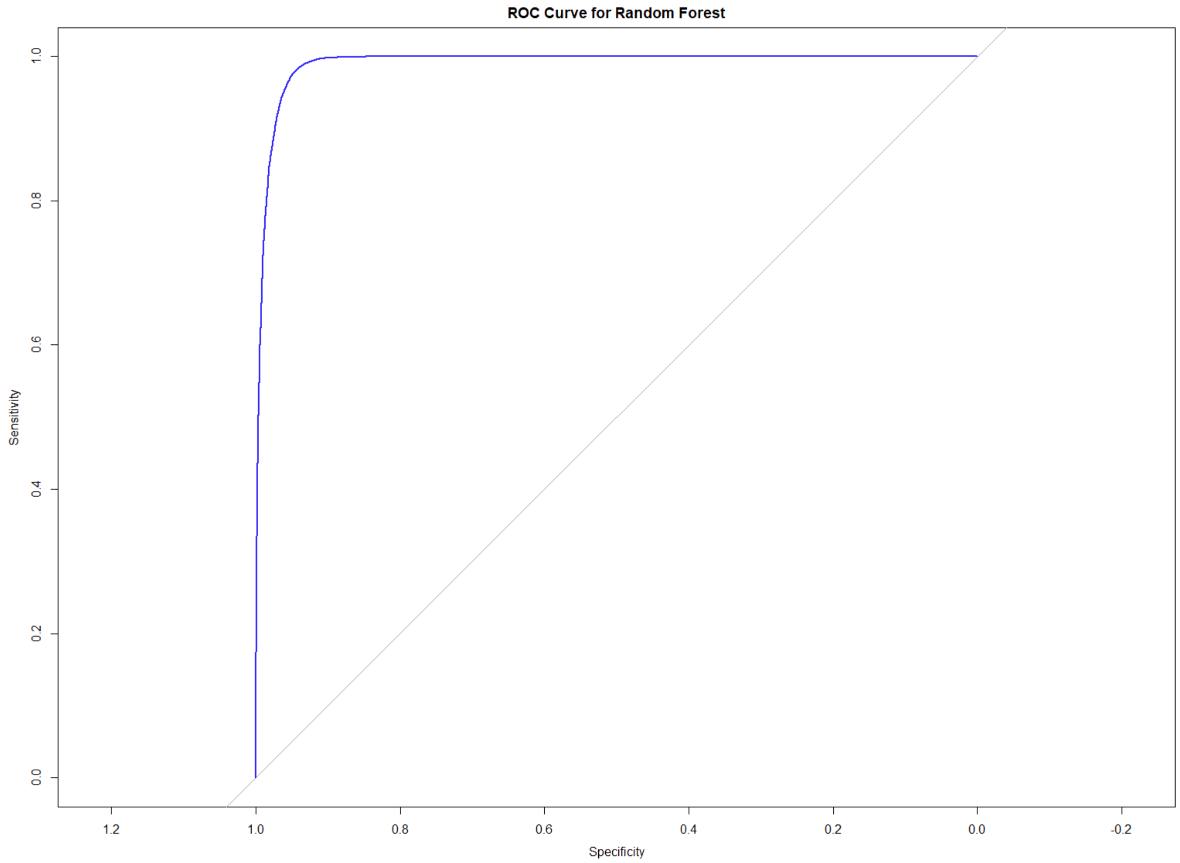
1. **Feature Importance Chart**:
   - Showcasing EXT_SOURCE_2, Debt-to-Income Ratio, and Employment Duration as top predictors.



Columns: Feature

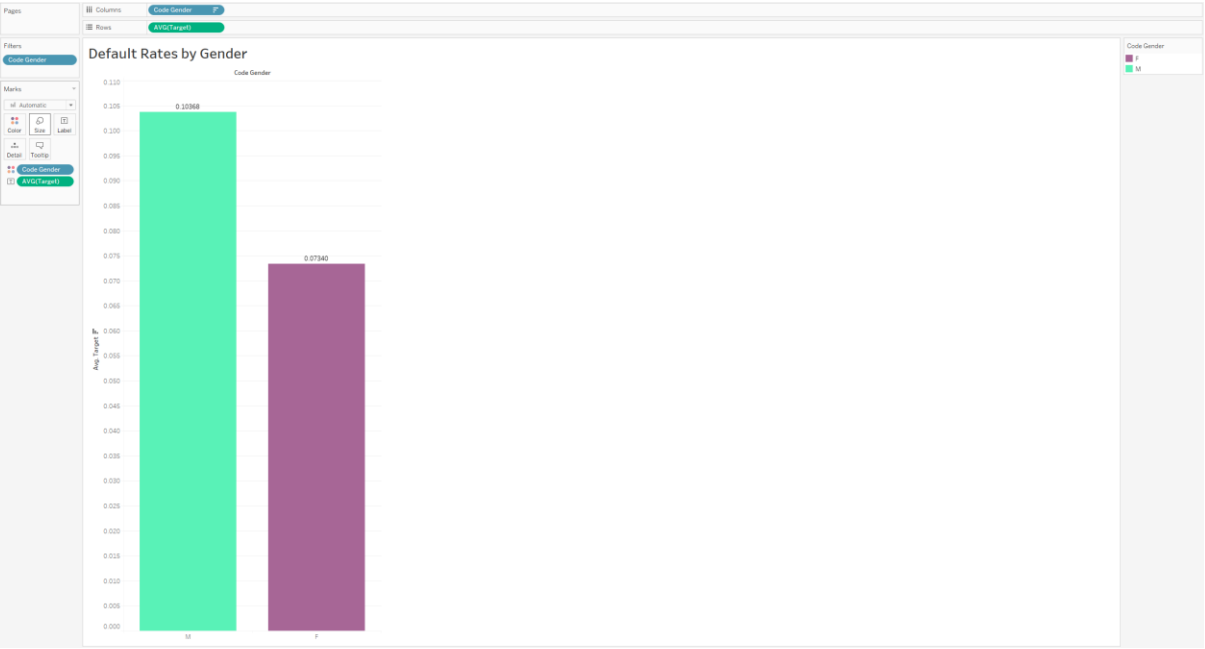
Rows: AVG(Importance)

2. **AUC-ROC Curve**:
   - Highlighting the exceptional discriminatory power of the Random Forest model.



3. **Default Rate by Gender**:

    - Bar chart indicating higher default rates among male borrowers.


Columns: Code Gender

Rows: AVG(Target)

Filters: Code Gender

Marks: Automatic

Color: Code Gender

Label: AVG(Target)

This case study demonstrates a robust application of data-driven approaches to solve real-world challenges, highlighting technical expertise and strategic thinking. Let’s connect to discuss how I can bring these skills to your organization!