Credit Risk Analysis & Prediction

Introduction:

Financial institutions are continuously faced with the challenge of evaluating the creditworthiness of loan applicants. Accurately assessing whether a borrower is likely to default or meet the required credit policies is crucial for minimizing financial risk and ensuring stable lending operations. In this project, the primary objective was to build a predictive model that classifies loan applicants based on their financial and credit-related attributes. The goal is to identify potential high-risk borrowers and assist financial analysts in making informed, reliable credit decisions.

Through exploratory data analysis (EDA) and machine learning techniques, the project uncovers key financial patterns, addresses data imbalances, and develops a classification model that supports more efficient decision-making. In particular, the focus is on understanding what factors influence whether a borrower meets the credit policy and identifying high-risk borrowers for the institution's risk management teams.

About the Data:

The dataset consists of various financial and personal attributes of borrowers, which are used to predict whether they meet the credit underwriting criteria (i.e., whether they are approved for a loan). Key features in the dataset include:

- **credit.policy:** A binary indicator where 1 means the borrower meets the credit policy, and 0 means they do not.
- **purpose:** The purpose of the loan (e.g., "credit_card", "debt_consolidation", "educational", etc.).
- **int.rate:** The interest rate of the loan, which is an indicator of the risk associated with the loan.
- **installment:** The monthly payment required if the loan is funded.
- **log.annual.inc:** The logarithm of the self-reported annual income of the borrower.
- **dti:** The debt-to-income ratio of the borrower.
- **fico:** The FICO credit score of the borrower.
- **days.with.cr.line:** The number of days the borrower has had a credit line.
- **revol.bal:** The borrower’s revolving balance at the end of the billing cycle.
- **revol.util:** The borrower’s revolving credit utilization ratio.

- **inq.last.6mths:** The number of credit inquiries in the last six months.
- **delinq.2yrs:** The number of times the borrower has been 30+ days overdue on a payment in the past two years.
- **pub.rec:** The number of derogatory public records (e.g., bankruptcy filings, tax liens, or judgments).

Problem Statement:

The project aims to answer the following business-critical questions:

1. **What financial factors (like interest rates, income, FICO scores) influence whether a borrower meets the credit policy criteria?**
    - **Answer:** Through analysis, it was found that features such as `inq.last.6mths`, `fico`, `days.with.cr.line`, and `revol.bal` have the highest feature importance and correlation to the credit policy. These features are pivotal in determining whether a borrower is approved for a loan.
2. **Is there an imbalance in the data regarding borrowers who meet vs. don't meet credit policies? How does this affect model performance?**
    - **Answer:** The dataset is imbalanced, with more borrowers meeting the credit policy. To address this, I used **SMOTE** (Synthetic Minority Oversampling Technique) to balance the data, ensuring that the model does not show bias towards the majority class.
3. **Which classification model performs best in predicting credit policy adherence: Logistic Regression, Random Forest, or others?**
    - **Answer:** Among various classification models, **Random Forest** performed the best in predicting whether a borrower meets the credit policy criteria.
4. **How well can the model predict risky borrowers?**
    - **Answer:** The Random Forest model was able to predict risky borrowers with an impressive **97% accuracy** on the test data, indicating that it is highly effective in classifying high-risk applicants.
5. **What are the most important features driving credit decisions, and how can financial institutions leverage them to reduce default risk?**
    - **Answer:** Features such as `inq.last.6mths`, `fico`, `days.with.cr.line`, and `revol.bal` are most strongly associated with loan approval. Financial institutions can focus on these factors when evaluating applicants to reduce default risk, for example, by monitoring applicants' recent credit inquiries or ensuring they have a healthy credit utilization rate.

Observations:

1. **Data Imbalance:**
    - There are more borrowers who meet the credit policy (1) than those who do not (0), which indicates an imbalance in the dataset. This can potentially skew the model's predictions if not handled properly.
2. **Interest Rates and Risk:**
    - Higher-risk borrowers, i.e., those who do not meet the credit policy, tend to have higher interest rates and lower credit scores (FICO scores). This aligns with the idea that riskier borrowers are charged higher rates to mitigate the lender's risk.
3. **Income Does Not Predict Credit Policy Adherence:**
    - Interestingly, annual income (`log.annual.inc`) is not a strong indicator of whether a borrower meets the credit policy. Customers who do not meet the credit policy come from a wide range of income levels, suggesting that other factors, such as credit behavior and debt management, play a more significant role in the credit approval process.
4. **Correlation between Credit History and Loan Approval:**
    - Borrowers with a longer credit history (i.e., higher `days.with.cr.line`) tend to be more likely to meet the credit policy. This suggests that borrowers with a stable credit history are seen as more reliable by the institution.

Conclusion:

This analysis not only uncovered the most critical factors influencing loan approval but also developed a high-performance classification model to predict whether a borrower meets the credit policy. By leveraging features such as FICO scores, recent credit inquiries, and revolving balances, financial institutions can make more informed decisions, mitigate risk, and enhance their lending strategies.

Key Business Insights:

- Financial institutions can focus on evaluating applicants' credit utilization, credit history length, and recent credit inquiries to better assess their risk.
- Addressing the imbalance in loan approval statuses through techniques like SMOTE can improve model performance, especially for predicting risky borrowers.
- The predictive model, which achieved a 97% accuracy rate, offers a reliable tool for loan approval decisions, helping companies make smarter, data-driven lending choices.

Next Steps for the Business:

- **Refining Credit Policies:** Based on the analysis, LendingClub can refine its credit policies by focusing more on high-impact features such as FICO scores, recent credit inquiries, and credit utilization.

- **Risk Management:** The business can further leverage this model to identify high-risk borrowers early in the loan application process, allowing them to adjust loan terms (such as interest rates) or reject risky applicants more effectively.

- **Real-time Decision Systems:** With further model optimization and real-time data input, this analysis could lead to the creation of a real-time loan approval system, improving operational efficiency and customer satisfaction.