Group E14

https://github.com/rauldsepp/datascienceproject

Team members

Mihkel Orasmäe

Imre Saks

Raul Raudsepp

Task 2

Identifying Your Business Goals

Background:

In the fast-paced financial sector, efficient loan processing is vital. Traditional manual evaluations are slow, error-prone, and often fail to capture full applicant profiles. Machine learning offers a way to analyze applicant data efficiently, speeding up decisions and identifying high-risk borrowers to reduce defaults and financial losses.

Business Goals:

- Automate Loan Evaluations: Build a machine learning model to predict loan approvals accurately.
- Improve Risk Management: Identify high-risk applicants to minimize defaults.
- Increase Efficiency: Streamline workflows, cut turnaround times, and reduce costs.
- Enhance Customer Experience: Deliver faster, more reliable decisions.
- Support Strategic Decisions: Use insights to refine lending policies and products.

Business Success Criteria:

- Achieve 90% accuracy in predictions.
- Keep false negatives and false positives below 5%.
- Reduce processing times by 50%.

Assessing Your Situation

Inventory of Resources:

- Human Resources: Three students
- Data: Kaggle datasets (`train.csv`, `test.csv`) with applicant and loan information. Loan approval dataset in kaggle.
- Tools: Python, libraries like pandas and TensorFlow, cloud-based computation, visualization tools (Matplotlib, Seaborn), GitHub, and project management platforms.

Requirements, Assumptions, and Constraints:

- Requirements: Meet the December 13th deadline and create an interpretable model.
- Assumptions: Datasets reflect real-world loan applicants and allow effective training.
- Constraints: Limited time and resources; potential data quality issues.

Risks and Contingencies:

- Data Quality: Handle missing or imbalanced data with cleaning and resampling.
- Model Performance: Test multiple algorithms and tune hyperparameters.
- Team Coordination: Set clear timelines and checkpoints to align progress.
- Technical Issues: Reserve time for debugging and seek support when needed.

Terminology:

- Loan Approval Prediction: Forecasting if a loan will be approved or rejected.
- Creditworthiness: A borrower's ability to repay loans.
- False Negative: Eligible applicants denied loans by the model.
- False Positive: Ineligible applicants approved by the model.
- Machine Learning Model: Algorithms that learn patterns to make decisions.

Costs and Benefits:

- Costs: Time for data preparation and model building; computational resources.
- Benefits: Faster loan decisions, reduced operational costs, improved risk management, and enhanced customer satisfaction.

Defining Your Data-Mining Goals

Data-Mining Goals:

- Build Models: Develop machine learning models to predict loan approvals.
- Feature Engineering: Identify and create impactful predictive features.
- Address Imbalance: Balance target classes to improve accuracy.
- Explain Decisions: Use techniques like SHAP values for transparency.

Success Criteria:

- Quantitative Metrics: Achieve 90% accuracy, high precision/recall, and AUC-ROC > 0.85.
- Qualitative Outcomes: Generate useful insights and align decisions with policies.
- Stakeholder Buy-In: Demonstrate effectiveness to users and decision-makers.

Task 3

Gathering Data

Defining Data Requirements:

The dataset must provide a comprehensive view of loan applicants, including personal demographics, financial standing, and loan details. The aim is to identify patterns linked to loan approvals and defaults by analyzing applicant profiles. Essential data fields include age, income, employment history, and loan-related attributes such as loan amount, interest rate, and repayment history.

Ensuring Data Availability:

The dataset includes a rich variety of information about applicants and their loan details. It is structured to facilitate meaningful analysis, containing sufficient records to evaluate the relationships between applicant attributes and loan outcomes. The available data ensures a robust foundation for pattern recognition and model training.

Defining Selection Criteria:

To ensure meaningful insights, the data must include diverse applicant profiles, loan purposes, and repayment histories. Selection criteria prioritize applicants with varying loan amounts, interest rates, and income levels, as well as different credit history lengths. This diversity enables a thorough analysis of patterns affecting loan outcomes.

Describing Data

The dataset consists of various fields, each playing a critical role in understanding loan outcomes:

1. id: A unique identifier for each record.
2. person_age: Applicant's age, segmented into categories to identify trends across age groups.
3. person_income: Applicant's income, categorized into ranges for comparative analysis.
4. person_home_ownership: Ownership status of the applicant's residence, such as RENT, MORTGAGE, or OWN.
5. person_emp_length: Length of employment, categorized by years, to understand its impact on loan approval.
6. loan_intent: Purpose of the loan, such as EDUCATION, MEDICAL, or PERSONAL, providing context for the application.
7. loan_grade: Credit grade of the loan, categorized as A, B, etc., indicating the applicant's creditworthiness.
8. loan_amnt: Loan amount requested, categorized to analyze trends in approval rates.
9. loan_int_rate: Interest rate for the loan, grouped into ranges for easier comparison.
10. loan_percent_income: The proportion of the loan amount to the applicant's income, highlighting affordability.
11. cb_person_default_on_file: Whether the applicant has a history of loan defaults (Yes/No).
12. cb_person_cred_hist_length: Length of the applicant's credit history, categorized for pattern identification.
13. loan_status: Target variable, indicating whether the loan was approved (binary values: 0 or 1).

Exploring Data

The exploratory phase involves analyzing distributions, correlations, and patterns in the dataset. Examples of analyses include:

- Age and Loan Status: Determine if specific age groups are more prone to loan defaults.
- Income and Loan Amount: Analyze whether higher incomes correlate with larger loan requests.
- Home Ownership and Defaults: Assess if homeowners are less likely to default compared to renters.
- Credit History and Loan Status: Investigate how the length of credit history impacts loan approval and default rates.

Preliminary exploration suggests a strong relationship between income levels and loan amounts, as well as between credit history length and loan outcomes. Analyzing these factors provides valuable insights into applicant behavior and risk profiles.

Verifying Data Quality

Completeness:
All essential fields, such as `loan_status` and `person_income`, are fully populated, ensuring no missing data that could compromise analysis.

Validity:
The dataset adheres to expected data types. For example, `loan_int_rate` is stored as a numeric value, and categorical fields like `person_home_ownership` contain valid labels. These checks confirm the dataset's reliability for modeling purposes.

Consistency:
Values across all fields are consistent with expected ranges. For example, age and income fall within plausible limits, while loan amounts and interest rates align with typical financial industry standards.

Accuracy:
The data has been cross-checked for anomalies and errors. Outliers, such as unusually high loan amounts or implausible employment lengths, have been identified and flagged for further investigation.

Task 4

Task 4: Planning Your Project (0.25 points)

Project Plan, we plan on completing the project together as a team:

1. Data Cleaning (ca 5 hours):
   During this step, the dataset will be prepared for analysis by handling missing values, correcting inconsistencies, and removing outliers. Key activities include standardizing numerical values, ensuring consistent formatting for categorical data, and normalizing features for better model performance.

2. Feature Engineering (ca 5 hours):
   New variables will be created to capture meaningful relationships in the data. Tasks include developing ratios (e.g., loan-to-income), encoding categorical attributes, and transforming existing features to enhance predictive capabilities.
3. Exploratory Data Analysis (ca 10 hours):
   The team will explore the dataset through visualizations and statistical summaries. This includes creating plots like bar charts, scatterplots, and heatmaps to identify trends, correlations, and anomalies, providing insights into how variables influence loan outcomes.

4. Model Development (ca 15 hours):
   Machine learning models will be implemented, trained, and optimized to predict loan approval. Algorithms such as logistic regression, decision trees, and ensemble methods will be evaluated using cross-validation, with a focus on maximizing accuracy and minimizing errors.
5. Evaluation and Poster Design (ca 5 hours):
   Model performance will be assessed using metrics such as precision, recall, and AUC-ROC. The final results will be compiled into a visually appealing poster that summarizes the project's methodology, findings, and impact.

Methods

1. Decision Trees: Chosen for their simplicity and interpretability, this method provides clear decision paths and is useful for identifying key decision criteria.
2. Random Forests: A robust algorithm used for its accuracy and stability, ideal for reducing overfitting and improving predictive performance.
3. Neural Networks: Applied to explore complex, non-linear relationships in the dataset, offering advanced modeling capabilities through deep learning techniques.

Tools

1. Jupyter Notebook: A flexible environment for coding, data analysis, and documenting results, providing an interactive workspace for the team.
2. Python: The primary programming language for the project, leveraging its libraries (e.g., pandas, scikit-learn) for data processing and modeling.
3. GitHub: A version control platform to manage code, track progress, and collaborate effectively within the team.