PROJECT REPORT

Project number: G11

Project title: SME Loan approval prediction

Team members: Romet Ustav, Maileen Rohtmets, Hendrik Treufeldt

Repository (GitHub)

Business understanding

Background
Banks and the SBA approve loans for small businesses; deciding an appropriate loan amount is time-consuming and can be inconsistent across officers.

Business goals

- Use historical SBA loan data to estimate how large a loan is reasonable for a new applicant.
- Support lenders and analysts by providing a data driven estimate of the loan amount a business is likely to be approved for.
- Improve consistency and speed of loan sizing decisions.

Business success criteria

- The model reduces manual analysis time per application.
- Predicted amounts are considered plausible by domain experts and reduce extreme outlier loan sizes.

Inventory of resources

- Historical SBA loan dataset.
- Python, pandas, scikit-learn, Jupyter, matplotlib, numpy…
- Skills in data cleaning, feature engineering, model training.

Requirements, assumptions, constraints

- The model must work on structured tabular data, with the features available at application time (no post approval fields).

- Assume relationships in historical data are roughly stable in the short term.
- Model must run in reasonable time on standard hardware and be explainable at a high level.

Risks and contingencies

- Data or policy drift may degrade performance → plan periodic retraining.
- Historical bias (by industry, region, etc.) may be learned → monitor and adjust if needed.
- Data quality: remaining errors or missingness in the data may degrade model performance; keep validation checks and logging in place.

Terminology

- Instance: one loan application (row).
- Features: input variables (NAICS, Term, NoEmp, etc.).
- Target: approved loan amount GrAppv.

Costs and benefits

- Costs: time spent on data preparation, modeling, evaluation, and deployment; computing resources; potential need for ongoing monitoring and retraining.
- Benefits: faster and more consistent loan sizing decisions, potential reduction in credit risk from extreme or inconsistent loan amounts, and better use of historical data in decision making.

Data-mining goals

- Build at least two predictive models that estimate GrAppv from the selected features.
- Achieve substantially better performance than a naive baseline that always predicts the mean or median loan amount (e.g., significantly higher R^2 and lower RMSE).
- Understand which features most strongly influence predicted loan amounts (e.g., via feature importance).

Data-mining success criteria

- Quantitative: On a held-out test set, the model achieves R^2 clearly above 0 (ideally > 0.5) and a low RMSE. The model is stable across different random train/test splits (performance does not vary wildly).
- Qualitative: Feature importance and partial patterns make domain sense (e.g., larger firms and certain industries tend to have higher predicted loan amounts). Loan officers (or reviewers) find the predicted ranges plausible and potentially useful as a decision support tool.

Data understanding

Gathering data

Outline data requirements

The goal is to predict the gross approved loan amount (GrAppv) for SBA-guaranteed loans from attributes known at application time, such as firm size, industry, geography, and loan characteristics. The data must therefore include historical loans with GrAppv and predictors like State, NAICS, ApprovalFY, Term, NoEmp, NewExist, CreateJob, RetainedJob, FranchiseCode, UrbanRural, RevLineCr, and LowDoc, all at the individual-loan level.

Verify data availability

The "SBAnational.csv" dataset was obtained as a CSV file with about 899k rows and 27 columns, one row = one loan. It contains all required fields plus additional columns (e.g., ChgOffDate, SBAAppv) that can be dropped because they are not available at decision time or are not of predictive value. The file loads successfully into pandas and can be processed on a standard laptop.

Define selection criteria

From the original 27 columns, only 12 predictor attributes plus the target GrAppv are retained: State, NAICS, ApprovalFY, Term, NoEmp, NewExist, CreateJob, RetainedJob, FranchiseCode, UrbanRural, RevLineCr, LowDoc, and GrAppv. Records with clearly invalid or missing critical values (e.g., missing NAICS, non-numeric GrAppv) are excluded to keep a high-quality modeling sample.

Describing data

After preparing the data, the working dataset contains 349641 loans and 88 fields + 1 target field. Each row represents one SBA-guaranteed loan. States are encoded as two-letter codes; NAICS is collapsed to its first two digits to represent broad industry sectors; ApprovalFY is the fiscal year of approval; Term is loan term in months; NoEmp, CreateJob, and RetainedJob are employee counts. NewExist, UrbanRural, RevLineCr, and LowDoc are

small categorical fields describing business age, location type, revolving credit, and documentation program, respectively. GrAppv is a positive, skewed numeric variable.

Exploring data

Initial exploration shows that most loans are relatively small: histograms of GrAppv reveal a strong skew with many loans between 50k and 300k and some larger loans. This suggests that a log transformation of the target may help some models. Year of approval is concentrated in the 1980s–2010s; NAICS sectors such as retail trade, accommodation and food services appear frequently. Employee counts and job creation/retention variables are also skewed with many small firms.

Verifying data quality

Systematic checks were performed for missing values, impossible values, and inconsistent coding. About 200k loans lacked a valid NAICS code; given the importance of sector, these rows were remove. Terms less than or equal to zero, retained jobs exceeding total employees, and invalid NewExist, UrbanRural, RevLineCr, or LowDoc codes were identified and filtered out. Categorical markers such as “N/A”, “null”, and empty strings were converted to missing before dropping. Numeric fields were cast to appropriate integer types, and NAICS was truncated to its two-digit sector code with a mapping to ensure only defined sectors remain. After these steps, remaining data showed no obvious quality issues and was judged suitable for modeling.

Planning the project

Task	Done by – hours worked
Preparing the data – removing inconsistent rows, missing/impossible values + dropping features that are not needed + encoding the data to a machine-readable format etc.	Romet – 25h
Writing code for the first model and training it, documenting and explaining the results, creating appropriate visuals to illustrate the process	Maileen Rohtmets – 15h
Optimizing the first model, hypertuning parameters, importance of features, finding the best model specific variables that produce more accurate results	Maileen Rohtmets – 15h

Writing code for the second model and training it, documenting and explaining the results, creating appropriate visuals to illustrate the process	Hendrik Treufeldt – 15h
Optimizing the second model, hypertuning parameters, importance of features, finding the best model specific variables that produce more accurate results	Hendrik Treufeldt – 15h
Creating a visually appealing, informative, concise vector poster with visuals illustrating our work	Romet Ustav – 5h