Mortgage Approval Prediction System

-Dhruv Prajapati
-Chetan Bhanushali

Introduction to Dataset

Home Mortgage Disclosure Act (HMDA) dataset contains detailed information about **mortgage applications** in the United States, including **loan characteristics, applicant demographics, property details,** and the outcomes of loan applications.

Original Dataset:

Size: 10 GB

100+ Columns

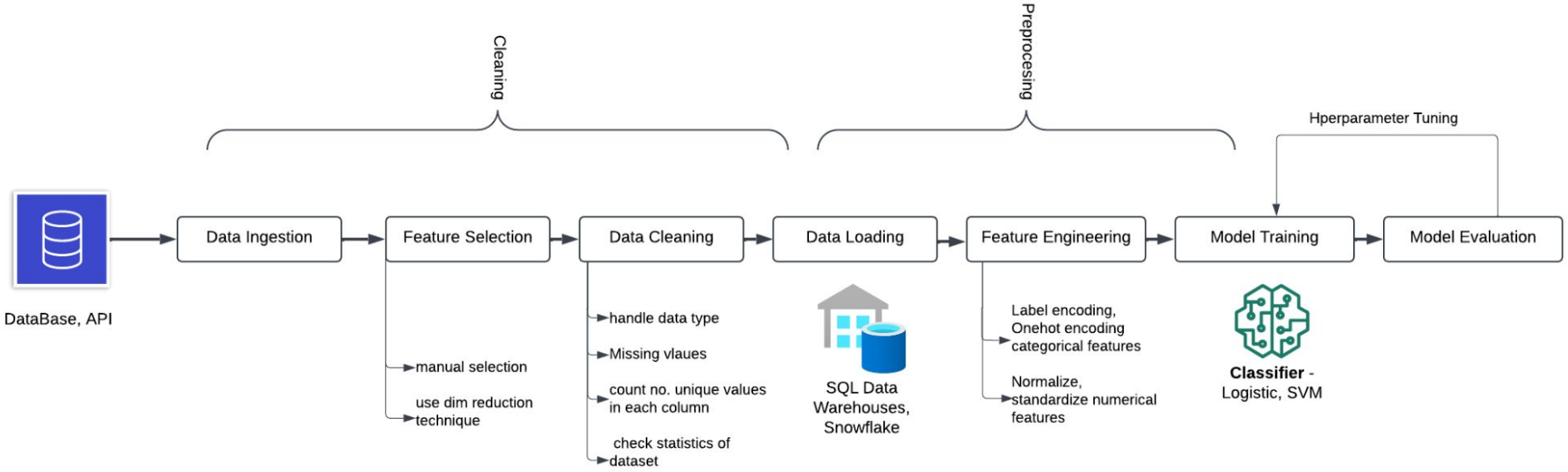
25M+ Rows

```
"lei": StringType(),
"loan_type": IntegerType(),
"loan_purpose": IntegerType(),
"loan_amount": FloatType(),
"interest_rate": FloatType(),
"loan_term": IntegerType(),
"action_taken": IntegerType(),
"income": FloatType(),
"applicant_age": StringType(),
"applicant_sex": IntegerType(),
"applicant_credit_score_type": IntegerType(),
"co_applicant_age": StringType(),
"co_applicant_credit_score_type": IntegerType(),
"derived_msa_md": IntegerType(),
"state_code": StringType(),
"county_code": StringType(),
"property_value": FloatType(),
"total_units": IntegerType(),
"occupancy_type": IntegerType(),
```

Objectives

- **Exploratory Data Analysis to find**
    - **Loan Approval Trends:** Identify patterns in approval/denial rates based on geography, lender, or applicant demographics.
    - **Borrower Demographics:** Investigate if certain groups have higher denial rates or access to less favorable loan terms.
    - **Geographic Trends**: Analyze loan distribution across states, counties.
- **Predict Loan application approval**
- **Estimate the interest rate of an application**

The Approach



Data Cleaning Techniques

Data cleaning ensures the dataset accuracy and reliability.

To ensure reliable analysis, we applied several data cleaning methods:

- Missing Values: - Identified and handled missing data
- Used imputation techniques to fill gaps
- Duplicates: - Removed duplicate records to avoid skewed results

Inconsistencies: - Standardized data formats and corrected inconsistencies

After data cleaning we are left with 20 columns and 19M rows.

Data Preprocessing Steps

Preprocessing prepares the data for model training:

- Data Normalization:
  - Scaled numeric features to a common range
  - Ensures fair comparison between features

- Encoding Categorical Variables:
  - Converted categorical data into numerical format
  - Used techniques like one-hot encoding

- Splitting the Dataset:
  - Divided data into training and testing sets
  - Training set: 80%, Testing set: 20%

Preprocessing is crucial for effective model training.

Feature Engineering Insights

Feature engineering enhances model accuracy:

- Generated new features from existing data
- Examples: Loan-to-Value ratio, Debt-to-Income ratio
- Improved model's ability to predict loan approval
- Identified key factors influencing loan decisions

Feature engineering adds valuable insights to the dataset.

Results

	Logistic	SVM	Factorization Machine	Decision Tree	Random Forest	Gradient Boosting
Accuracy	0.7063	0.7097	0.5579	0.9248	0.7304	0.9461
Precision	0.6549	0.7773	0.6491	0.9309	0.8034	0.9461
Recall	0.7063	0.7097	0.5779	0.9248	0.7304	0.9461
F1-score	0.6266	0.5933	0.5773	0.9214	0.6382	0.9453
ROC	0.6873	0.6058	0.5814	0.8819	0.8198	0.9744

Model Prediction and Results

Logistic Regression

One of the models used for prediction

Decision Trees

Another model used for prediction

Random Forest

Model with the highest accuracy

Accuracy, Precision, Recall, F1-Score

Performance metrics used to evaluate models

Conclusion and Implications

Key Findings

- Data cleaning and preprocessing are essential for reliable analysis

- Feature engineering significantly improves model accuracy

- Random Forest model is effective in predicting loan approval

Implications for Lenders

- Better understanding of factors influencing loan approval

- Informed decision-making in mortgage lending

Future Studies

- Explore additional features and models

- Continuous improvement of prediction accuracy

Our analysis provides valuable insights for the mortgage lending industry.