Documentation: Loan Approval Prediction using ML

Contents

1 Overview

This documentation provides a detailed guide on building and understanding the loan approval prediction model. Two distinct approaches are used: a **Self Organizing Map (SOM)** and a **Classification-based model**. These models aim to classify loan applications based on customer data, predicting whether a loan will be approved or not.

2 Problem Statement

Loan approval is a critical process in the financial sector. Banks and financial institutions assess multiple factors before approving loans to minimize risk. The challenge is to build models that can automate and improve the accuracy of loan approval decisions by predicting approval based on customer features like income, employment status, and credit history.

3 Dataset Description

The dataset contains 690 records with 16 features. Each record represents a customer's demographic and financial information. The features are categorized as follows:

3.1 Features

- **CustomerID**: Unique identifier for each customer.
- **Gender**: Encoded as 0 for female and 1 for male.
- **Marital Status**: Encoded as 0 for single and 1 for married.
- **Employment Status**: Encoded as 0 for unemployed and 1 for employed.
- **Job Security**: Encoded as 0 for insecure and 1 for secure.
- **Credit History**: Encoded as 0 for bad and 1 for good credit history.
- **Personal Loan**: Encoded as 0 for no personal loan and 1 for existing personal loan.
- **Age**: Numerical value representing the customer's age.
- **Income(LPA)**: Income in Lakhs Per Annum.
- **Current Loan Amount**: Total amount of existing loans.
- **Number of Dependents**: Number of dependents.
- **Number of Existing Loans**: Number of existing loans held by the customer.
- **Years at Current Residence**: Number of years the customer has lived at their current residence.
- **Current Bank Balance**: The current balance in the customer's bank account.
- **Collateral Value**: Value of the collateral provided for the loan.
- **Class**: Target variable indicating loan approval status (1 for approval, 0 for disapproval).

3.2 Target Variable

Class: Binary variable (0 = Loan not approved, 1 = Loan approved).

4 Methodology

4.1 Approach 1: Self Organizing Map (SOM)

Self Organizing Maps are a type of unsupervised neural network used for clustering. They are particularly good at identifying patterns in high-dimensional datasets. In this project, SOMs are used to group similar customer profiles based on their demographic and financial features to detect potential approval groups.

Steps:

1. Load the dataset.
2. Preprocess the features (`X`) by normalizing the values for better clustering.
3. Apply the Self Organizing Map algorithm to map customer profiles.
4. Visualize the clusters to see distinct customer segments that may have similar approval likelihoods.

Output: SOM provides clusters of customers, helping to understand which customers are more likely to get approved based on their feature similarity to others.

4.2 Approach 2: Classification Model

A classification model is used to predict whether a loan will be approved or not, based on the given features. The process involves:

Steps:

1. **Data Preparation**: The dataset is loaded, features (`X`) and the target (`y`) are separated.
2. **Preprocessing**: Handle missing values, scale numerical features, and encode categorical features.
3. **Model Selection**: Choose a suitable classification algorithm such as Logistic Regression, Decision Tree, Random Forest, or others.
4. **Training the Model**: Fit the model using training data.
5. **Evaluation**: Evaluate the model using performance metrics like accuracy, precision, recall, and F1-score.
6. **Prediction**: Predict loan approval status for new applicants.

Output: A model capable of predicting the loan approval status based on input customer features.

5 Model Training and Evaluation

5.1 SOM Model

Training: The SOM model is trained using an unsupervised learning algorithm. No labels are used during training.

Evaluation: After training, the SOM generates a 2D grid where each neuron represents a cluster of customer data points. By visualizing the SOM, clusters of customers can be identified.

5.2 Classification Model

Training: The classification model is trained using the labeled dataset, where features predict whether the loan was approved (`Class`).

Evaluation Metrics:

- **Accuracy**: Measures how often the classifier is correct.
- **Precision**: Measures how many selected items are relevant.
- **Recall**: Measures how many relevant items are selected.
- **F1-Score**: Harmonic mean of precision and recall, providing a balanced evaluation metric.

6 Code Snippets

6.1 Data Loading and Preprocessing

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

# Load dataset
dataset = pd.read_csv('Loan_Approval_Dataset.csv')

# Separating features and target variable
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values

# Splitting into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

6.2 SOM Model Example

```
from minisom import MiniSom

# Initializing SOM with dimensions 10x10
som = MiniSom(x=10, y=10, input_len=X_train.shape[1], sigma=1.0, learning_rate=0.5)

# Training SOM with the customer data
som.random_weights_init(X_train)
som.train_random(data=X_train, num_iteration=100)

# Visualizing the SOM clusters
plt.bone()
plt.show()
```

6.3 Classification Model Example

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Initializing and training the Random Forest classifier
classifier = RandomForestClassifier(n_estimators=100, random_state=0)
classifier.fit(X_train, y_train)

# Making predictions
y_pred = classifier.predict(X_test)

# Evaluating the model
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification␣Report:\n", classification_report(y_test, y_pred))
```

7 Results

7.1 SOM Model

Clusters of customer profiles were generated, showing distinct groups that may have similar loan approval statuses. These clusters help understand patterns in customer demographics and financials.

7.2 Classification Model

The classification model predicts loan approval status with a specific accuracy, providing a data-driven approach to automating loan approval decisions.

8 Conclusion

This project demonstrates two approaches for predicting loan approval:

- **Self Organizing Map (SOM)** for clustering customer profiles.
- **Classification-based models** for supervised prediction of loan approval.

These models can help financial institutions make faster and more accurate loan decisions by analyzing customer data and identifying key approval patterns. Further improvements could include advanced feature engineering and hyperparameter tuning for better model performance.

9 Future Work

- **Hyperparameter Tuning**: Optimize model parameters for better performance.
- **Feature Engineering**: Derive new features from existing data to improve prediction accuracy.
- **Model Deployment**: Deploy the classification model as a web service for real-time loan approval prediction.

This documentation outlines the steps, methods, and results for developing a loan approval system. It serves as a comprehensive guide for replicating the model and understanding its application in the financial sector.