Lab 1 – Report

By Jayasinghe C.H – 210247B

Link to the GitHub Repo containing the notebook →
Machine_Learning_Stuff/lab_1 at main · Chamikajaya/Machine_Learning_Stuff (github.com)

Content

Introduction

1.1 Background and Overview

Feature engineering, as a practice, is pivotal in the development of predictive models. By intelligently selecting and transforming variables, one can extract meaningful patterns from raw data. This report focuses on applying feature engineering and data preprocessing techniques to a loan default prediction dataset obtained from a finance company in the United States.

The dataset comprises information about previous loan applicants, including details such as credit scores, financing queries, and address information. The overarching objective is to discern patterns indicative of potential loan defaults, empowering stakeholders to make informed decisions such as denying loans, adjusting loan amounts, or setting interest rates based on the perceived risk.

1.2 Motive and Objective

The primary objective behind this lab assignment develops a machine learning classification model capable of accurately predicting loan defaults. To achieve this, the following tasks was undertaken:

- **Feature Selection / Removal:** Employing data cleaning and feature scoring techniques to identify and retain pertinent features.
- **Feature Encoding:** Converting categorical variables into a format interpretable by machine learning algorithms.
- **Feature Engineering Techniques:** Implementing methods such as one-hot encoding, ordinal encoding, and date encoding to capture nuanced information within the dataset.
- **Dimensionality Reduction:** Leveraging techniques like SelectKBest to focus on the most impactful features.

- **Model Development:** Utilizing the XGBoost Classifier for creating a predictive model and iterating on its performance through validation.
- **Shapely Values for Explainable AI:** Employing SHAP analysis to interpret and provide transparency to the developed model.

Data Overview

2.1 Dataset Description

The dataset under consideration originates from a finance company in the United States and revolves around the prediction of loan defaults. It encompasses a wealth of information regarding past loan applicants, providing insights into their financial profiles and behaviors. The dataset consists of approximately 860,000 observations, each characterized by 150 variables, including both features and the target variable, 'loan_status.'

Key Features include:

- Credit scores
- Quantity of financing queries
- Address details (zip codes, states)
- Collections information
- Various financial and personal factors

These features are crucial in identifying patterns that may indicate whether a person is likely to default on a loan. The dataset is rich but poses challenges such as missing values, outliers, and a mix of data types. To facilitate understanding, a comprehensive data dictionary (**DataDictionary.xlsx**) was provided to us.

2.2 File Details

The dataset is distributed across multiple files, each serving a specific purpose in the overall machine learning workflow:

- **DataDictionary.xlsx:** This file contains essential information about the dataset, offering a comprehensive guide to the various features and their meanings.
- **train.csv:** This file constitutes the training data, encompassing all variables (features) and the target column ('loan_status'). It is utilized for training the XGBoost Classifier, a fundamental step in developing the predictive model.
- **valid.csv:** This file serves as the validation data, mirroring the structure of the training data. It includes features and the target column ('loan_status') and is used to validate and assess the performance of the machine learning model trained using the training data.
- **X_test.csv:** This file represents the test data, featuring all available variables (features) but without the target column ('loan_status'). It is employed to predict unknown observations using the developed model.

Feature Engineering

Feature engineering is a critical phase in the predictive modeling pipeline, involving the exploration of relationships between different features to derive valuable insights and optimize predictive performance.

Here are some of the feature engineering and feature selection techniques that were used by me while developing the model.

3.1 Feature Selection / Removal

To enhance computational efficiency and model interpretability, a subset of features was removed based on domain knowledge and correlation analysis. The correlation matrix, a numerical representation of feature

relationships ranging from -1 to 1, played a pivotal role in identifying highly correlated features.

3.2 Feature Encoding

To facilitate machine learning model training, categorical variables underwent effective encoding. Label encoding and one-hot encoding techniques were employed for categorical features, capturing essential information about loan characteristics and purposes.

3.3 Advanced Feature Engineering Techniques

Several advanced techniques were employed to extract nuanced information from specific columns, ensuring a richer representation of the dataset:

- **Term Extraction**: The 'term' column was transformed to extract numerical values, aiding in numerical representation.
- **Grade to Integer Conversion**: 'sub_grade' underwent ordinal encoding, capturing the ordinal nature of loan grades.
- **Numeric Part Extraction**: Functions like 'convert_to_int' facilitated the extraction of numerical components from string columns.
- **Ordinal Encoding of Employment Length**: 'emp_length' was ordinal encoded to capture the progressive nature of employment duration.

3.4 Dimensionality Reduction

Dimensionality reduction was crucial for model efficiency and interpretability. Leveraging the correlation matrix and SelectKBest, the following techniques were applied:

- **Correlation Analysis**: The correlation matrix was instrumental in identifying highly correlated features. Features exhibiting a correlation coefficient above a predefined threshold, such as 0.85, were considered for removal to reduce redundancy. By leveraging

the correlation matrix, this feature engineering process aimed to streamline the feature space, improve computational efficiency, and enhance model interpretability. The identification and removal of highly correlated features contributed to a more concise and efficient dataset, reducing the risk of overfitting and laying the foundation for a robust loan default prediction model. ((*Refer the Jupiter notebook for the correlation matrix)*)

- **SelectKBest**: Utilizing the chi-squared scoring function, SelectKBest focused on the most relevant features. Iterative optimization of the number of features ('k') aimed to improve model accuracy.

3.5 Imputation and Scaling

To handle missing values, **SimpleImputer** imputed with the mean. **MinMaxScaler** was employed to ensure all features had non-negative values, optimizing the dataset for machine learning algorithms.

Model Development

4.1 Data Splitting

The dataset underwent a division into training and validation sets, employing the *train_test_split* function from scikit-learn. This segregation is vital to train the model on one subset and assess its performance on another, ensuring a robust evaluation of the model's generalization capabilities.

4.2 XGBoost Classifier

The XGBoost Classifier, was chosen as the model for this endeavor. The model was initialized and trained on the selected features, harnessing the power of gradient boosting to achieve superior predictive accuracy.

4.3 Model Evaluation

The performance evaluation of the XGBoost Classifier took place on the validation set. The assessment involved computing accuracy scores to gauge the model's adeptness at generalizing to unseen data, providing valuable insights into its reliability.

4.4 Iterative Feature Engineering

The model development process embraced an iterative feature engineering approach. The SelectKBest algorithm, coupled with chi-squared scoring, played a pivotal role in selecting the most influential features. The optimization of the number of features ('k') was a strategic move aimed at augmenting model accuracy.

4.5 Model Prediction on Test Data

The final iteration of the model was applied to the test data after undergoing preprocessing and feature engineering. Predictions were generated using the selected features, and the results were systematically stored in a CSV file which is already submitted to the moodle.

4.6 Model Performance:

The XGBoost Classifier exhibited outstanding accuracy on both the validation and test datasets, showcasing its robust generalization capabilities. The accuracy scores are as follows:

- **Accuracy on Test Data:** 99.55%
- **Accuracy on Validation Data:** 99.52%

Conclusion & Challenges faced

This project has successfully employed feature engineering and precise data preprocessing techniques to boost the predictive capabilities of a loan default classification model. The XGBoost Classifier emerged as a robust choice, showcasing remarkable accuracy during evaluation on both validation and test datasets.

Challenges Faced:

Navigating challenges in handling missing values, optimizing categorical variable encoding, and addressing potential overfitting from highly correlated features marked significant hurdles. Striking a balance between model accuracy and generalization proved to be an difficult task. Despite these challenges, the project's success underscores the pivotal role of strategic feature engineering in augmenting the efficacy of loan default prediction models.