Loan Approval Prediction Model

Abel Kebekabe • Alexander Bradley • Sebhat Gezehey

Introduction

The Loan Approval Prediction Model is a machine learning model developed to predict whether a loan application should be approved or not based on a range of factors such as gender, marital status, education, number of dependents, income, loan amount, credit history, and others. This model was developed using the Amazon Web Services (AWS) Sage Maker Notebook environment and deployed to a Sage Maker endpoint for easy access.

System Architecture and Design

The flow diagram below summarizes the design and architecture of the implemented system. A developer authenticates into AWS as an IAM user. The input training data set is stored in an AWS S3 bucket. The model was developed in an AWS SageMaker Notebook environment which is a fully managed Amazon Elastic Compute (EC2) instance. The Jupyter notebook code creates a SageMaker training job. The training job reads input data from S3 and after training completes, SageMaker saves the resulting model artifacts to an Amazon S3. SageMaker also stores model images in a private repository in Amazon ECR. The code also sets up a SageMaker deployment, endpoint configurations and deploys the model to a SageMaker endpoint.
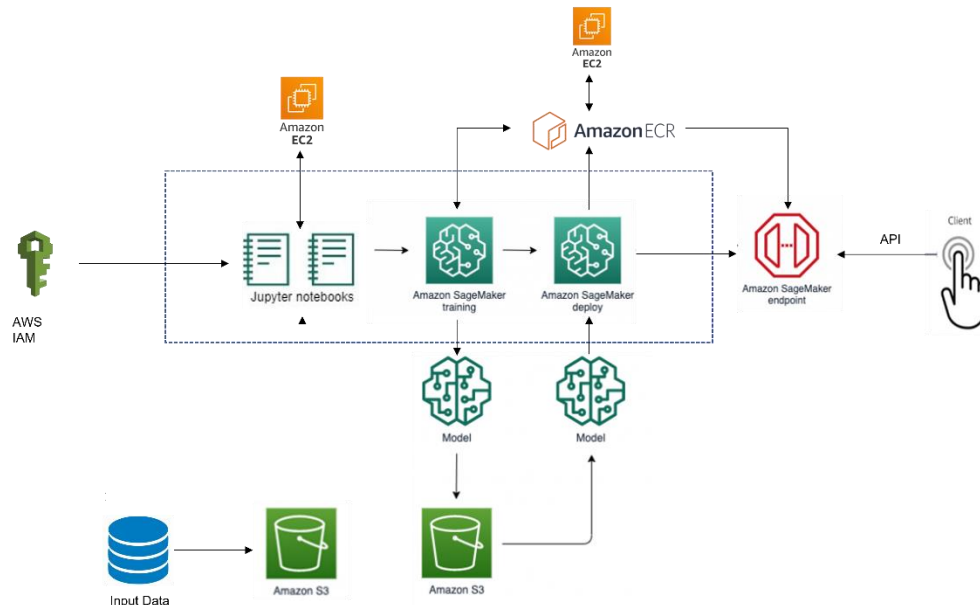
Data Collection and Preparation

The dataset used for training the model was obtained from Kaggle.com and contained information about loan applicants, including their gender, marital status, education, number of dependents, income, loan amount, credit history, and other factors. Please see reference for a link to the data set. The dataset was cleaned and preprocessed to ensure that it was suitable for training the machine learning model. This involved removing missing values, encoding categorical variables, and splitting the dataset into training and validation sets.

Model Development

The Loan Approval Prediction Model was developed using a supervised machine learning algorithm, specifically a binary classification algorithm, trained on the cleaned and preprocessed dataset. The model was trained to predict whether a loan application should be approved or not based on the input features. Several machine learning algorithms were tested during the development of the model, including random forests, Weighted Ensemble, eXtreme gradient boosting (XGBoost). A Bayesian hyperparameter optimization technique was used to find the best values for the hyperparameters. The best performing algorithm was selected based on its performance on the validation set and ease of deployment.

Version	Model	Description
Baseline	Amazon SageMaker AutoML	AutoML was used to produce a quick analysis of multiple models. WeightedEnsemble was the best model.
Version 1.0	Random Forest	Model was chosen for being a viable solution to a binary classification problem.
Version 1.1	XGBoost	Chosen as it was one of the better performing algorithms from the AutoML.
Version 1.2	XGBoost	Additional tuning of XGBoost was done.

Model Evaluation

The performance of the Loan Approval Prediction Model was evaluated using several metrics, including accuracy, precision, recall, and F1-score. In the context of loan approval, both false positive and false negatives have similar costs. Since the F1 score combines the precision and recall scores, we used the F1 score as the objective metrics. In addition, F1 score is a good metric to use for evaluating binary classification with an imbalanced target variable.

The figure below summarizes the performance of different versions of the test models against a baseline model. Version 1.1 had the highest F1 score but it was determined that the model was overfitting. Further optimization of version 1.1 resulted in version 1.2 eliminating some of the overfitting observed in version 1.1. Version 1.0 and version 1.2 were the candidates for production based on F1 score. Although version 1.0 had a better F1 score than version 1.2, we chose version 1.2 as the production model for ease of deployment on a SageMaker endpoint. The production version achieved an F1 score of 0.82 on the validation set, which indicated that it was performing well in predicting whether a loan application should be approved or not.

Version	Model	F1	Accuracy	Precision	Recall	AUC
Baseline	Amazon SageMaker AutoML	0.677	0.839	0.913	0.538	0.247
Version 1.0	Random Forest	0.85	0.77	0.77	0.88	0.67
Version 1.1	XGBoost	0.96	0.96	0.96 (1.0 on Approval)	0.96	0.942
Production Version 1.2	XGBoost	0.82	0.84	0.85	0.84	0.748

Deployment

The Loan Approval Prediction Model was deployed to an AWS SageMaker endpoint for easy access. This allows a lending institution to easily input loan application information and receive a prediction on whether the loan should be approved or not. The model endpoint was also configured to handle multiple requests at the same time, ensuring that it could handle a high volume of loan applications.

Test

To ensure optimal performance of the model, the following tests were done:

- Test the preprocessing steps and ensure that the data is transformed correctly.
- Test that the model loads and runs without error
- Test the model outputs and ensure that expected results are obtained
- Test that the model endpoint can be invoked and returns a response
- Test the model on a smaller data set of test data to ensure it is working as expected

Overall, testing was done to ensure that the loan approval machine learning model deployed on SageMaker endpoint is accurate, reliable and meets expected performance expectations.

Conclusion

The Loan Approval Prediction Model developed on AWS SageMaker Notebook and deployed to a SageMaker endpoint is an effective machine learning model for predicting loan approval based on a range of factors such as gender, marital status, education, number of dependents, income, loan amount, credit history, and others. The model achieved an F1 score of 0.82 on the validation set and was deployed to a SageMaker endpoint for easy access.

Reference

1. https://www.kaggle.com/datasets/mirzahasnine/loan-data-set
2. Prepare data for machine learning - Amazon SageMaker Data Wrangler - Amazon Web Services
3. Machine Learning with Amazon SageMaker :: Amazon SageMaker Workshop (sagemaker-workshop.com)