Proposal

Domain Background

One of the ongoing areas of interest within Finance domain is to mitigate the risks involved in consumer lending business while ensuring good customers are onboarded with a smooth customer experience while applying for loans and/or credit cards.

Inorder to mitigate the risk in this area, customers are evaluated based on several parameters like earning potential, credit score, spending patterns etc. to identify the ones likely to setlle their payments on time and avoid any loss for the financial institution lending them with financial help.

It is equally important for the financial institutions to identify customers likely to default in their payments and take appropriate measusres like rejecting their credit request or providing them with less risky alternatives like credit cards with lower credit limits.

Several companies already have employed different machine learning models like logistic regression and decision trees to facilitate the process of identifying target customers for their lending business. However, given the competitive landscape there is a need to keep exploring new factors that can improve the predictions. Listing some prominent examples below:
https://www.sciencedirect.com/science/article/pii/S2214785321035148
https://dl.acm.org/doi/10.1145/3411501.3419431
https://rstudio-pubs-static.s3.amazonaws.com/209149_8de3e66249f442288ed51b07fd384c12.html#/

Given my 9+ years of experience in the FinTech domain, this problem falls in my area of interest and I am interested in exploring various Machine learning techinques to identify dependable and robust solutions for this problem.

Problem Statement

Few months back, American Express (Amex), which is the largest payment card issuer globally, launched 'American Express - Default Prediction' challenge on Kaggle. The objective is to classify the customers either as a potential non-defaulter (0) or a potential defaulter (1) by analyzing their monthly customer profile. The customer is expected to pay the due amount in 120 days after their latest statement date failing which they are considered to be a defaulter.

This is a binary classification, supervised learning problem for which several machine learning techniques can be explored.

Following is the link to the said featured competition posted by Amex: https://www.kaggle.com/competitions/amex-default-prediction/overview

Dataset and Inputs

The data set provided in the kaggle challenge is a huge dataset containing 5.53 million records and data for 458913 customers in the training dataset. The test dataset contains another 11.4 million records and data for 924621 customers. And both datasets have close to 190 feature columns.

The training labels have majority of good customers (close to 75%) and a relatively smaller percentage of defaulters (25%). Thus, the labels are highly imbalanced and this aspect needs to be considered while evaluating the model.

The input features are anonymized and normalized and are broadly classified into below categories as described in the data page of the challenge:

- D_* = Delinquency variables
- S_* = Spend variables
- P_* = Payment variables
- B_* = Balance variables
- R_* = Risk variables

Given that the huge training dataset size (16.39 GB), a lighter version of the training dataset (1.64 GB) in parquet format obtained from below link will be used for the purpose of capstone project: https://www.kaggle.com/datasets/raddar/amex-data-integer-dtypes-parquet-format

For further datasize reduction, the training data will be subsampled to 1% of the record size i.e. close to 100K records instead of 5 million records in the original dataset.

Solution Statement

The goal is to try to build a machine learning model that captures the potential defaulters with high probability while maintaining a low false positive rate i.e. avoiding misrepresenting a good customer as a potential defaulter.

To achieve this various machine learning models like Logistic regression, Xgboost, and other ensemble models will be explored.

Benchmark model

An initial model for the binary classification problem will be created using logistic regression to get a baseline performance on the dataset. The performance of the model will be accesed based on the chosen evaluation metric.

Later the same dataset sample used for training the earlier model will be used to train models using Autoglunon's tabular predictor and their performance will be compared to that of the baseline model created with traditional logistic regression approach.

Evaluation Metrics

The evaluation metric will be chosen based on how the data is subsampled. If the target class imbalance is reduced in training dataset then accuracy, ROC_AUC etc. will be used as the objective metric. If subsampling preserves the imbalance then metrics like F1 score, Sensitivity etc. will be considered.

Project Design

The following steps will be followed to build the desired model:

1. Exploratory Data Analysis: An initial exploratory data analysis will be performed to understand the datatype of the features, the data in each feature, the correlations that exist. Based on this understanding relevant columns will be retained as it is, or imputed for missing values, or deleted if it has little information. Dimensionality reduction will be performed if applicable and then data will be subsampled to training, validation and test sets.

2. Evaluation metric selection: An appropriate evaluation metric will be selected based on the distribution of majority and minority classes in the training and validation dataset.

3. Identifying the best predictor: AutoGlunon and Hyperparameter tuning will be used to identify good models. Measures will be adopted to avoid overfitting to the training data.

4. Model testing: The best model identified will be used to evaluate the model's performance on test data.

5. Deployment and Inference: Model will be deployed on Sagemaker and a lambda function will be created to drive inference from the model.