AMERICAN EXPRESS DEAFULT PREDICTION

Course:

- Advance Data Science/Architecture

Professor:

- Ram Hariharan

Group Members:

- Bhagyashree Nair (002139969)
- Sathwik Hegde (002194830)
- Chandana Uddehal (002105599)
- Prathik Gawand(002102812)

Background:

American Express is a globally integrated payments company. The largest payment card issuer in the world, they provide customers with access to products, insights, and experiences that enrich lives and build business success.

Whether out at a restaurant or buying tickets to a concert, modern life counts on the convenience of a credit card to make daily purchases. It saves us from carrying large amounts of cash and also can advance a full purchase that can be paid overtime.

How do card issuers know we'll pay back what we charge? That's a complex problem with more potential improvements, to be explored.

Credit default prediction is central to managing risk in a consumer lending business. Credit default prediction allows lenders to optimize lending decisions, which leads to a better customer experience and sound business economics.

In this project, we leverage an industrial scale data set to analyze various factors and build an effective model using machine learning techniques to predict weather the customers will be able to pay back the lenders.

Objective:

The objective is to predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile.

Dataset Key Specifications:

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

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

Working:

1. We performed data preprocessing on this dataset- Null value handling: Data imputation was performed by Mean, Median and Mode to analyze the effect of each imputation method.
2. We performed Exploratory Data Analysis of the profile variables- Delinquency, Spend, Payment, Balance, Risk. We did a Distribution analysis of all the variables to understand if it's skewed or normally distributed. Most variables showed a skewed distribution. We plotted a correlation of the independent features with the target.
3. Feature Engineering: We performed Dimensionality Reduction using Principal component Analysis.
4. We implemented 4 models:
    - Logistic regression
    - XGBoost
    - RandomForestClassifier
    - Support Vector
5. Hyperparameter Tuning for was performed for each model along with 3 techniques implemented for XGBoost
    - XGBoost Classifier with GridSearch
    - XGBoost with RandomizedSearchCV
    - XGBoost with BayesSearchCV
    - RandomForestClassifier using GridSearchCV
6. We analyzed the accuracies of each model- with and without hyper parameter tuning to understand which model works best for our dataset. Summarized visualization plotted to shows the results.

Results and Conclusion:

- Logistic Regression, after hyperparameter tuning, gave an accuracy of 87.21%

- XGBoost gave an accuracy of 88.62%
  After hyperparameter tuning:
  GridSearch: 88.62%
  RandomizedSearch: 89.25%
  BayesSearch: 90.02%

- RandomForestClassifier gave an accuracy of 90.46%
  After hyperparameter tuning:
  GridSearch: 90.25%
  This was the only case to record a lower accuracy (0.2%) after hyper parameter tuning

- Support Vector after hyperparameter tuning, gave an accuracy of 86.44%

- After a comparative study on the models above, we conclude that the classification model using RandomForestClassifier is the best model with an accuracy score of 90.46% without and 90.25 % with hyperparameter tuning.

References:

- https://www.kaggle.com/code/girishkumarsahu/american-express-default-prediction-ml-model
- https://www.kaggle.com/code/devsubhash/amex-eda-default-prediction
- https://www.kaggle.com/code/jinweitu/amex-default-prediction-with-lgbm-model-comparison
- https://towardsdatascience.com/hyperparameter-tuning-in-python-21a76794a1f7