**Done by: Mohammad Masarra, Aziz Lahiani**

**Part 2 (AI Project Technical Methodologies)**

This report documents the development of a machine learning model to predict customer churn based on transactional and demographic data. The project includes data preparation, feature engineering, model training, evaluation, inference, and model tracking using MLflow. The primary goal is to predict which customers are likely to churn, enabling targeted retention strategies.

**Data Preparation**

The data was read from a 'Commerce Data.xlsx' file and underwent initial preprocessing to handle null values and ensure data consistency. Summary statistics and correlations were computed to gain insight into the data distribution and relationships between features.

**Feature Engineering**

Numerical and categorical features were identified for use in the model. To address multicollinearity, highly correlated features like CouponUsed and OrderCount were noted. Feature scaling was performed using StandardScaler to normalize numerical features, and categorical features were encoded using one-hot encoding.

**Model Training and Selection**

A range of models, including Logistic Regression and Random Forest, were trained on the processed dataset. The models were evaluated based on accuracy, with Random Forest achieving the best score. This model was selected for its higher performance and absence of overfitting problems. Hyperparameter tuning was performed using GridSearchCV, but encountered shape mismatch errors during the process.

**Model Evaluation**

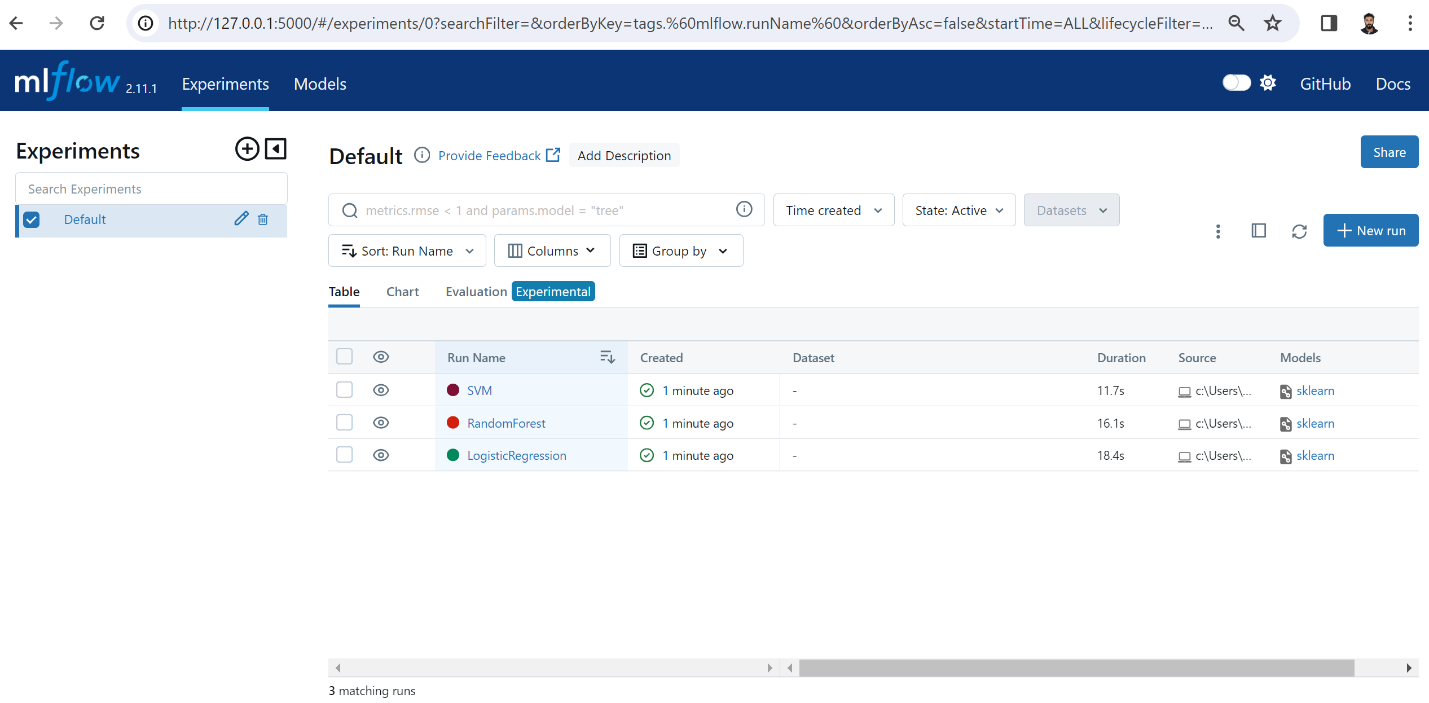
The models' performance was measured using accuracy as the metric. Random Forest achieved the highest accuracy and was subsequently chosen for making predictions on new data. The models were saved using joblib for persistence.

**Model Inference**

The trained Random Forest model was used to make predictions on unseen data. Predictions were combined with customer IDs to provide actionable insights for the business.

**Model Tracking**

MLflow was used to track experiments, parameters, and metrics, allowing for organized and reproducible model development. The test accuracy was logged, and the best parameters were identified for future reference.



**Conclusion**

The churn prediction model demonstrates good predictive performance with an accuracy rate exceeding 80%. Feature importance analysis with SHAP values adds transparency to the model's decisions, which is crucial for trust and adoption in business contexts. The model can inform customer retention strategies and is tracked for continuous improvement and reproducibility.

**Cloud Deployment**

We propose Google Cloud Platform, where datasets and machine learning models are stored in Google Cloud Storage for secure and easy access. The development and training of the model are facilitated by AI Platform Notebooks, which offers a managed JupyterLab environment with comprehensive ML framework support. For deployment, AI Platform Predictions is used to serve predictions from the trained model, leveraging google cloud platform managed services for scalable, efficient, and effortless model serving without the need for infrastructure management. This end-to-end cloud solution ensures a robust, scalable, and seamless operation from model development to prediction serving.

**Part 3: AI project Technical Methodologies**

**Waterfall Plot for a Specific Data Set Point**

The waterfall plot details how each feature's value influences a particular prediction relative to the base value (the average model output across the dataset). In the provided visualization, features such as Tenure and CashbackAmount increase the probability of the event we're predicting, whereas OrderAmountHikeFromLastYear and City\_Tier decrease it. This plot is a granular look at the decision-making process for a single data point, highlighting the push and pull of individual features.

**Force Plot for a Specific Data Set Point**

The force plot for the individual prediction demonstrates that OrderAmountHikeFromLastYear has a significant negative impact on the outcome, strongly suggesting a lower likelihood of the event being predicted. In contrast, Tenure and CashbackAmount provide positive contributions, indicating that longer tenure and higher cashback amounts are associated with an increased likelihood of the event. The red and blue shadings reflect the push-pull dynamic between the features, with the resulting prediction leaning slightly towards a higher probability due to the strong influence of Tenure.

**Summary Plot for the Whole Dataset**

The summary plot provides a macro-level view of feature importances across the entire dataset. It showcases which features have the most significant impact on model output. For example, Tenure and DaySinceLastOrder appear as the most influential features, suggesting that the length of customer engagement and recent activity are strong indicators in the model's predictions. This global perspective is useful for identifying which factors are consistently important in predicting outcomes.

**Bee Swarm Plot**

The bee swarm plot displays the distribution of the SHAP values for each feature across all data points. It allows us to see not just the impact of each feature but also the variability of this impact. Features with a wide spread of SHAP values, such as Tenure, indicate a varied influence on the model output across different observations. In contrast, a tight cluster suggests consistent behavior across the dataset.

**Dependence Plot**

The dependence plot reveals the relationship between a feature's value and its SHAP value, often indicating interaction effects. For instance, as Tenure increases, the impact on the prediction fluctuates, suggesting that other variables may interact with Tenure to influence the prediction.

**Interpretation of Results**

The SHAP analysis clearly indicates that certain features are pivotal in influencing the model's predictions. Longer tenure and more recent transactions are associated with a lower likelihood of churn, while increases in the order amount from last year and living in certain city tiers contribute to a higher likelihood of churn. The variability of impact across different customers, as shown in the bee swarm plot, emphasizes the complexity of the model's decision process and the individualized nature of the predictions.

**References**

* Shap Documentation : <https://shap.readthedocs.io/en/latest/>
* Shap with python

<https://www.youtube.com/watch?v=MQ6fFDwjuco&list=PLqDyyww9y-1SJgMw92x90qPYpHgahDLIK>

* Mlflow Documentation

<https://mlflow.org/docs/latest/index.html>

* Random Forest Classifier

<https://stackoverflow.com/questions/46137945/random-forest-classifier>

* GridSearchCV for multiple ML Models

<https://www.youtube.com/watch?v=MQ6fFDwjuco&list=PLqDyyww9y-1SJgMw92x90qPYpHgahDLIK>

* ChatGPT for different errors faced during doing this project