**BAN6800: Business Analytics Capstone**

**Module 4 Assignment: Business Analytics Model**

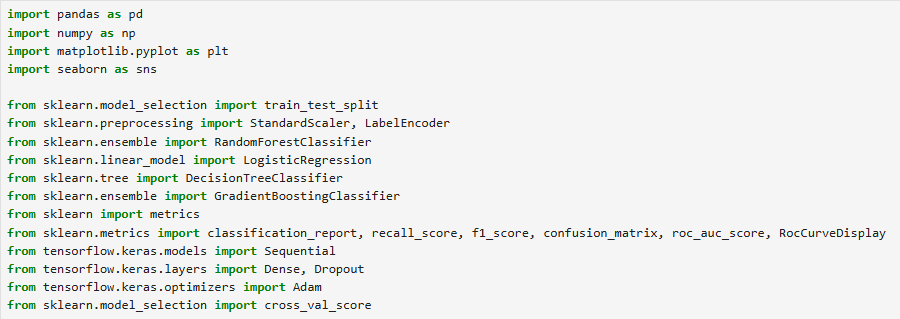
**Name:** Voke H. Edafejimue **| Learner ID:** 143304

**Taco-Tel’s Shield360 Project – Model Implementation Plan**

**Project Objective**

The Shield360 project seeks to build a business analytics model to pinpoint subscribers likely to abandon their subscription. Armed with this information, the company can take immediate action on retention strategies to endear these customers to the brand. This project will help Taco-Tel proactively target high churn risk subscribers with ads that will improve the customer lifetime value.

**Step 1: Import the necessary libraries**



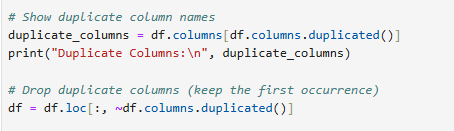
**Step 2: Data Preprocessing and Cleansing**

Our data was previously cleaned; however, we needed to ensure data quality, remove inconsequential columns to prevent leakage, and encode categorical data (Selvaraj, 2024).

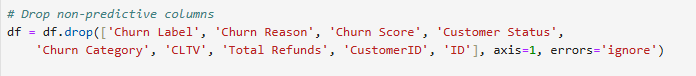
First, the cleaned data was loaded.



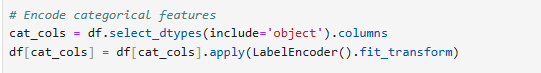
Next, show and remove duplicate columns.



Also, all identifiers, non-predictive data, and high-leakage columns, considered irrelevant to the business problem, were removed.

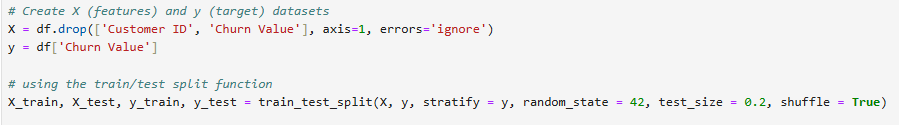


Furthermore, we encoded all categorical data.

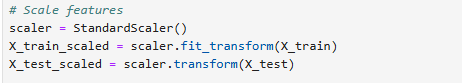


**Step 3: Feature Engineering**

First, the data was divided into two parts: 80% training data and 20% testing data. This ratio was chosen using stratified sampling to mitigate class imbalances.



The data was normalized using the Standard Scaler.



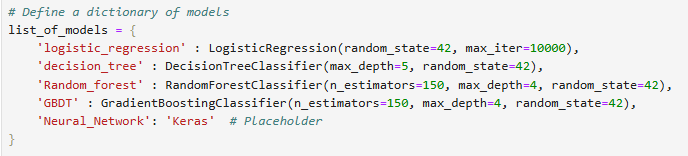
**Step 4: Model Development and Training**

We chose several models for this project. Each model was trained using 80% of the cleaned data, and the same tests were used for comparison (Singh, 2024).

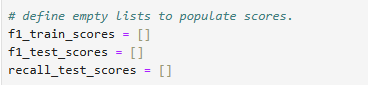
Models chosen:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosted Decision Trees (GBDT)
5. Multilayer Perceptron (MLP) using Keras – a neural network model

First, a dictionary of all the models is defined.



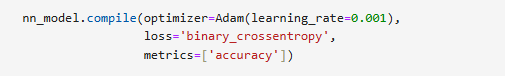
Also, empty lists were initialized to capture the relevant scores as we loop through the various models.



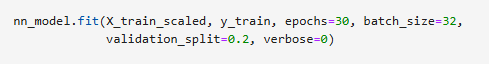
Neural networks work differently from other models; therefore, we check to see if the model type is Neural Network to build the model with multiple layers.



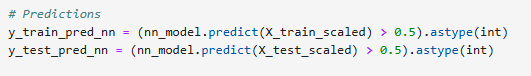
The model is then compiled using binary cross-entropy and optimized with ‘Adam’.



The model is then trained using 30 epochs with 32 samples.



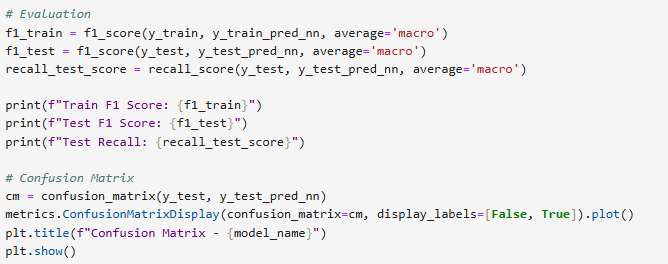
Predictions are then made – Churn = 1 or No churn = 0.



The model is cross-validated.



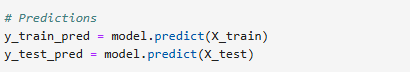
The model is evaluated using F1 scores, recall, and the confusion matrix.



If the model selected from the dictionary is not a neural network, it follows a separate process. The already defined and initialized model is trained to understand the relationship between the data features and the target labels.



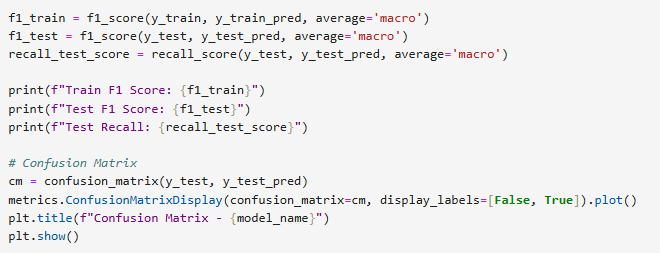
Predictions are made on both the training and testing datasets.



The model is cross-validated.



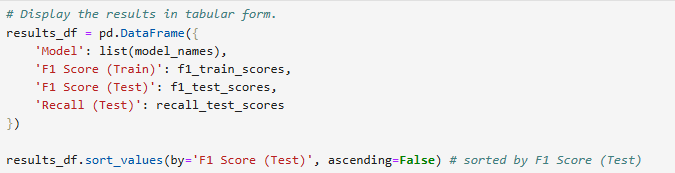
The models are graded on performance using the F1 score, recall, and confusion matrix to determine the best model.

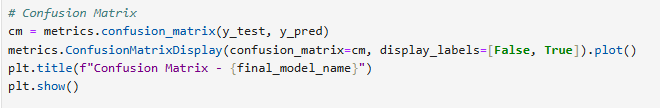


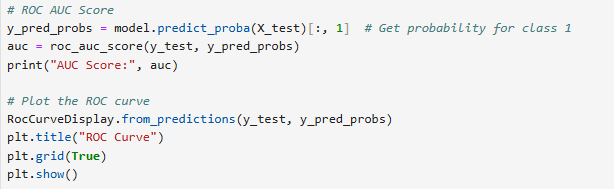
**Step 5: Model Validation and Evaluation**

The following metrics have been chosen to evaluate all 5 models for comparison on performance.

1. F1 Score – matches precision and recall scores
2. Recall Score – weighs the model’s response to True churn situations.
3. The Confusion Matrix – displays a visual representation of the prediction outcomes showing True Positives, True Negatives, False Positives, and False Negatives.
4. ROC AUC – Receiver Operating Characteristics Area Under the Curve - a graph for plotting True Positive Churn cases against False Positive Churn cases (Petriconi, 2021).







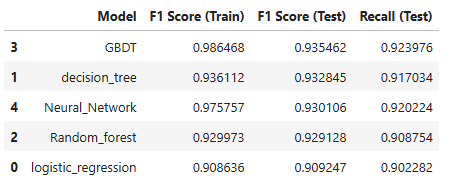
**Step 6: Model Acceptance Criteria**

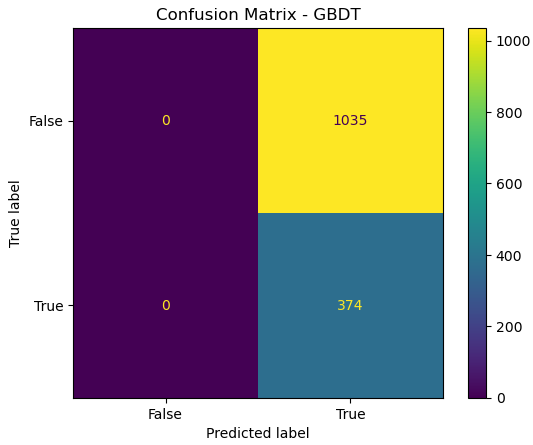
The chosen model must have a minimum performance score to align with the business and project goals.

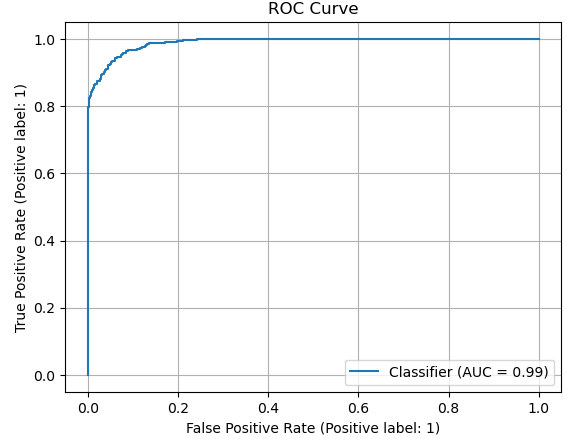
1. F1 Score >= 0.90, that is, 90%
2. Recall Score >= 0.90, that is, 90%
3. The train and test evaluation scores should be within a 5 to 10% difference. This will help mitigate overfitting.

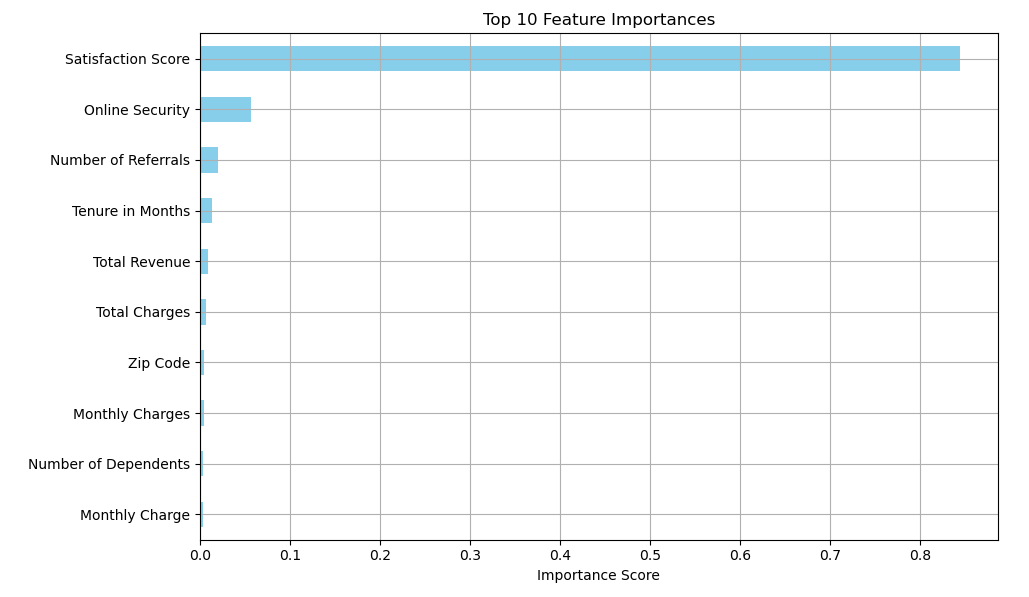
**Step 6: Analyze the Results**

The scores are sorted by the F1 Score (Test) to aid the best model selection.









*Interpretation*

The GBDT model outperforms the other models on the training and test datasets. The Multilayer Perceptron (MLP), Decision Tree, and Random Forest models performed well too. Though considered the simplest and easiest to understand, the Logistic Regression model had the lowest F1 score.

Also, an AUC score of 0.99 suggests a 99% chance that the GBDT model will correctly identify a subscriber with high churn risk. The solution is reliable and can be used to carry out subscriber retention strategies.

Features like the 'Satisfaction score' prove to be the most important factor in the dataset. Other features also play vital roles in the chosen models (Adiaturb, 2023)

**Conclusion**

Taco-Tel’s Shield360 project has developed multiple churn prediction models using a cleaned dataset that addresses the business question: Why do subscribers churn? After the dataset was cleansed, five models were developed, trained, tested, and evaluated. These models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Decision Trees (GBDT), and a Neural Network - Multilayer Perceptron (MLP).

GBDT performed the best with an F1 score of 0.935 and a recall of 0.924, meaning that it has a stronger predictive capability and balances the False Positives and False Negatives well. Employing model evaluation techniques like the confusion matrix, recall, AUC-ROC, and the F1 score, the project has met its target: a smart churn prediction model.

*Recommendations*

* The Best Model - deploy the GBDT model as its predictions are superior in accuracy and generalizability to other models.
* Use the selected model for churn prevention. Prioritize contacting customers with high churn risk with targeted campaigns.
* Encourage subscribers to opt for 1-to-2-year contracts, as these customers are less likely to abandon their subscription.
* Upgrade the company’s tech support to improve response turnaround time and offer self-service. A significant number of churned subscribers have unresolved tech complaints.
* Focus on value-added bundle services like cloud storage, gadget protection, and streaming services.
* Review the charges for subscribers with high bills and low usage.
* Fine Tune and Iterate – continuously monitor the evaluation metrics for changes and fine-tune the model.
* Further enhancements
* Use ‘GridSerachCV’ or ‘KerasTuner’ for hyperparameter tuning.
* Implement a real-time scoring pipeline with ‘Flask’ or ‘FastAPI’

**References**

* Natassha Selvaraj. February 1, 2024. How to Build a Customer Churn Prediction Model in Python? 365 Data Science. <https://365datascience.com/tutorials/python-tutorials/how-to-build-a-customer-churn-prediction-model-in-python/>
* Himanshi Singh. November 18, 2024. 10 Techniques to Solve Imbalanced Classes in Machine Learning. Analytics Vidhya. <https://www.analyticsvidhya.com/articles/class-imbalance-in-machine-learning/>
* Adiaturb. June 6, 2023. Customer Churn Prediction with Python — End-to-End Machine Learning Project. Medium. <https://medium.com/@adiaturb/customer-churn-prediction-with-python-end-to-end-machine-learning-project-93ccc6b2218f>
* Luca Petriconi. August 11, 2021. Churn Modeling: A detailed step-by-step Guide in Python. Medium. <https://medium.com/@lucapetriconi/churn-modeling-a-detailed-step-by-step-guide-in-python-1e96d51c7523>