**Homework 2: Kaggle Competition - Quitting Smoking Prediction**

**Team Information**

* **Kaggle Group Name:** Adir Serruya
* **Participants:** Adir Serruya, 316472455

**Overview**

**Objective**

The objective of this project was to develop a machine learning model to predict the likelihood of quitting smoking based on bio-signal features. The competition metric was the **Area Under the Curve (AUC)**, and the goal was to maximize this score.

**Stage A: AutoML Implementation**

**AutoML Packages Used**

**1. TPOT -** [**https://epistasislab.github.io/tpot/**](https://epistasislab.github.io/tpot/)

* **Methodology:**  
  TPOT is an AutoML tool that automates the process of machine learning model selection and hyperparameter tuning using genetic programming. The dataset was provided to TPOT, and the tool automatically explored pipelines for preprocessing, feature selection, and model building. The optimized pipeline was exported for further evaluation.
* **Hyperparameters Tuned:**
  + Population size: 10
  + Generations: 10
  + Mutation rate: 0.9
  + Crossover rate: 0.1
* **Results:**
  + AUC on Train Set: 0.989
  + AUC on Public Leaderboard: 0.848

### **2. AutoGluon -** <https://auto.gluon.ai/dev/tutorials/tabular/index.html>

* **Methodology:**  
  AutoGluon is an AutoML framework that automates model selection, hyperparameter optimization, and ensembling. The framework was applied to the dataset in its tabular mode. It trained multiple models (e.g., LightGBM, Random Forest, Neural Networks) and created an ensemble of the best-performing models.
* **Hyperparameters Tuned:**
  + Time limit for training: Default
  + Preset configurations: medium quality preset
* **Results:**
  + AUC on Train Set: 0.805
  + AUC on Public Leaderboard: 0.775

### **Stage B: Custom Model Development**

#### **Best Performing Algorithm:**

* **Algorithm Name:** XGBoost

#### **Preprocessing**

* **Steps Taken:**
  1. **Feature Engineering:** Applied two custom functions (create\_extra\_features and generate\_features) to create additional features, including interaction features, polynomial features, ratio features, log transformations, statistical aggregations, and health condition flags.
  2. **Feature Scaling:** Scaled the dataset using **Min-Max Scaler** to normalize feature ranges.
  3. **Variance Thresholding:** Removed features with variance below a threshold of 0.01 to reduce dimensionality and eliminate low-variance features.

#### **Model(s) Used**

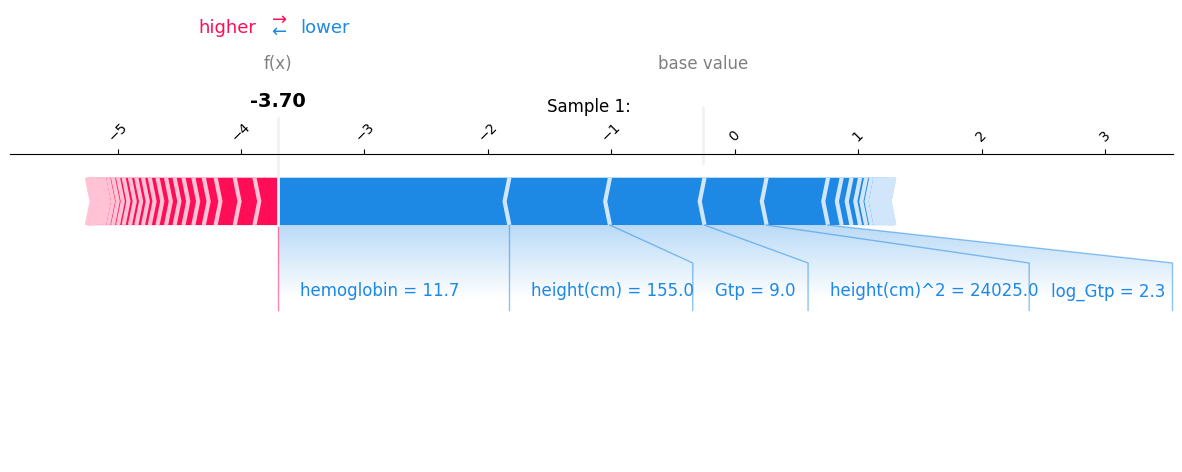
* **Algorithm Name:** XGBoost
* **Key Hyperparameters Tuned:**
  + n\_estimators: Tuned between 50 and 1000
  + max\_depth: Tuned between 3 and
  + learning\_rate: Tuned logarithmically between 0.01 and 0.3
  + subsample: Tuned between 0.5 and 1.0
  + colsample\_bytree: Tuned between 0.5 and 1.0
  + colsample\_bylevel: Tuned between 0.5 and 1.0
  + reg\_lambda: Tuned logarithmically between 1e-3 and 10.
  + reg\_alpha: Tuned logarithmically between 1e-3 and 10.
* **Ensemble Techniques:** No ensemble methods were used in this iteration, but hyperparameter optimization was conducted using **Optuna**, which ensured the best model configuration.

#### **Results**

* **AUC on Train Set:** 0.8902
* **AUC on Public Leaderboard:** 0.87306

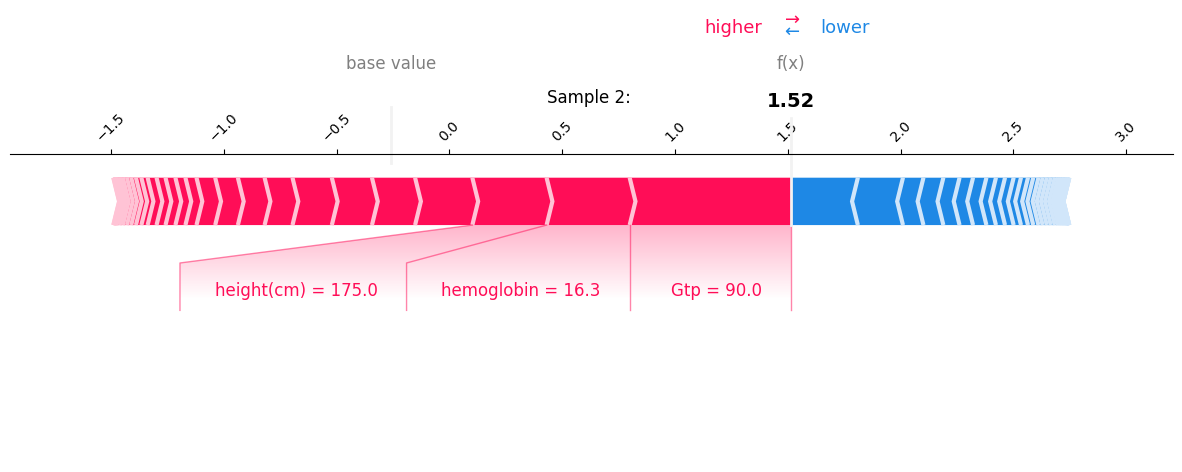
**SHAP Analysis**

**SHAP Explanations for 3 Selected Samples**

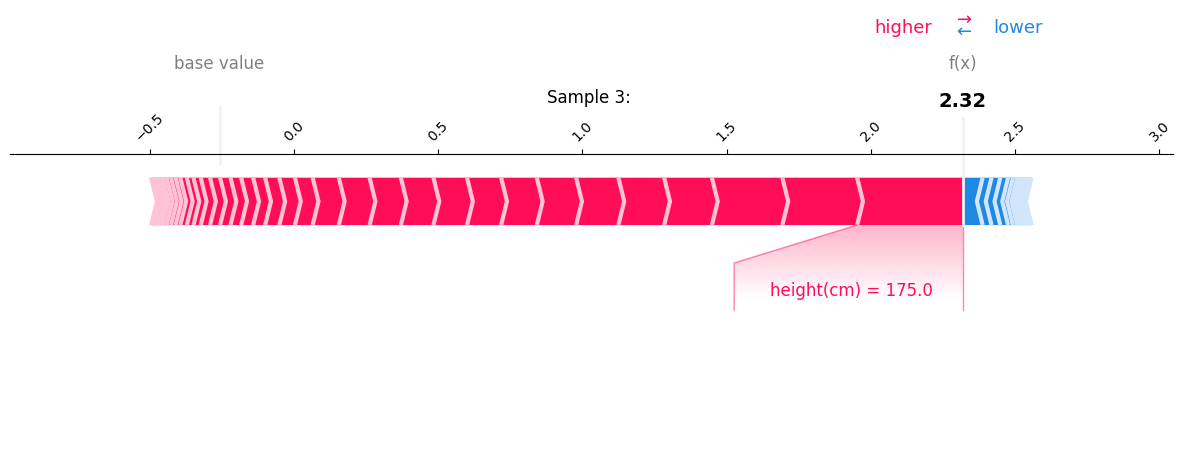
1. **Sample 1**

* Explanation: For the above sample we see that the values fro hemoglobing, height, Gtp and their various transfromations lowered the probability for a smoker.

Probability calculation example: in this example. (i.e. the probability for a smoker is around 2.4%)

1. **Sample 2**

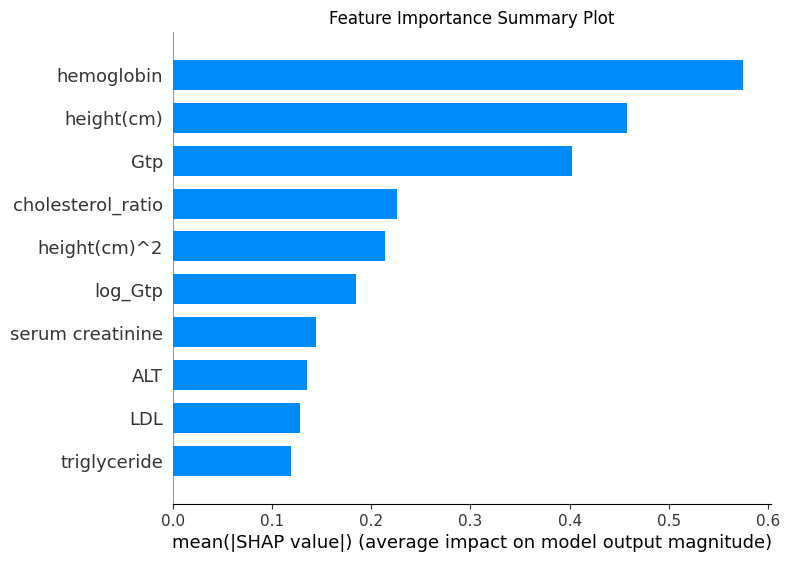
* Explanation: For the above sample we see that the values from hemoglobing, height, Gtp increased the probability for a smoker.

1. **Sample 3**

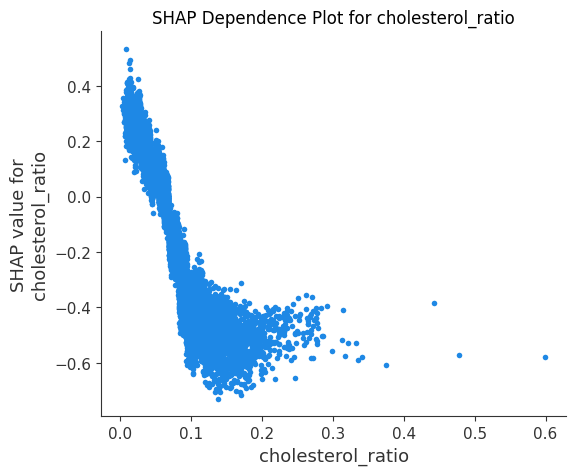
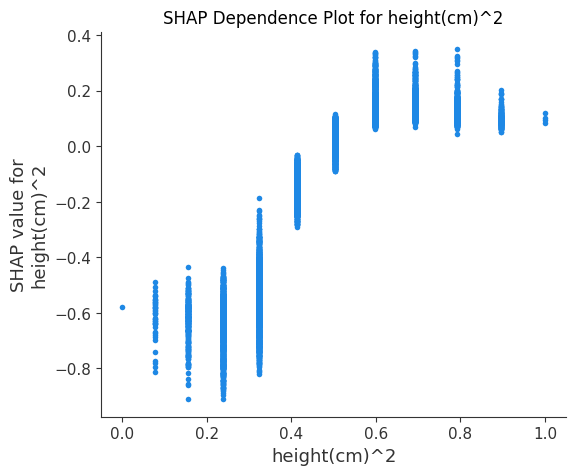
* Explanation: For the above sample we see that the values from height increased the probability for a smoker. ( If I were to guess why, probably because being 1.75m tall = high probability for a male = increased probability for a smoker )

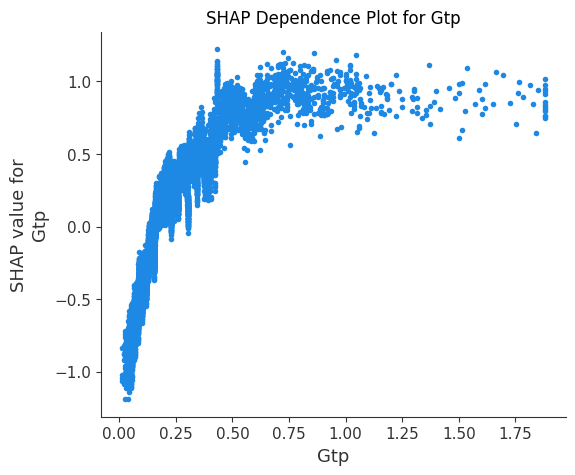
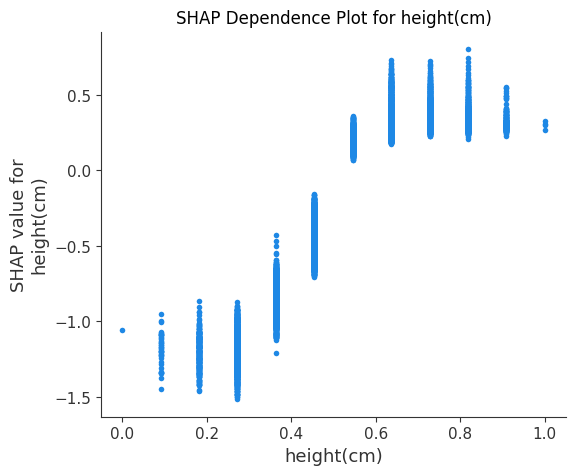
**Feature Importance Analysis**

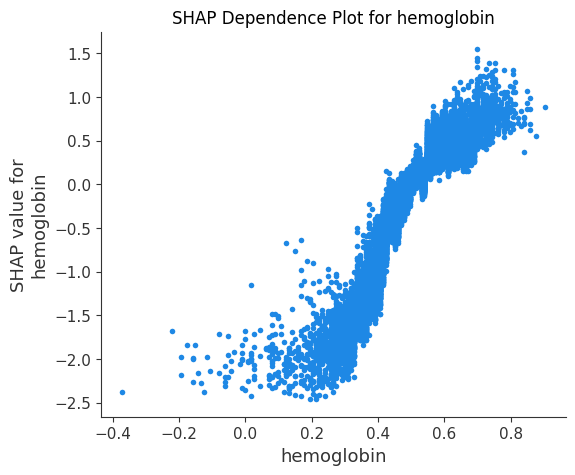
* **Summary Plot**



* **Analysis:** Recent research highlights the relationship between hemoglobin, triglycerides, and cholesterol levels with smoking behavior. **Height, in my view, likely acts as a proxy for gender**, as smoking rates are significantly higher among males compared to females.
* **Top 5 Feature Dependence Contribution Plots**

****

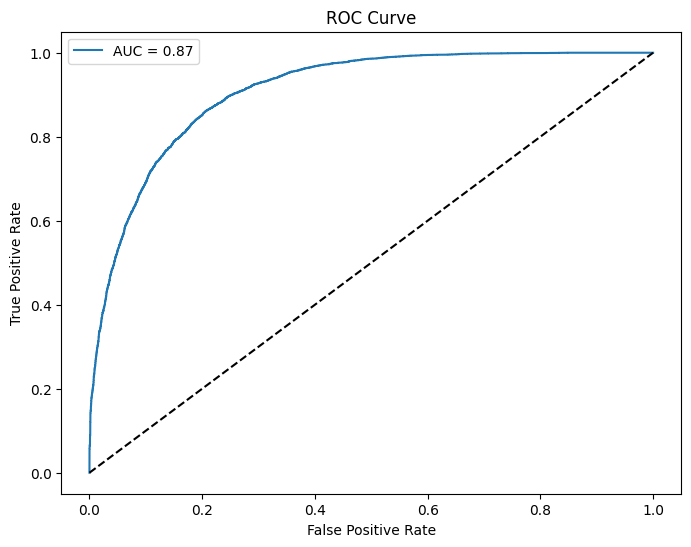
****

****

**Submission Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Run #** | **Method** | **Key HP** | **AUC(Train)** | **AUC(Public)** |
| **1** | **TPOT** | **CV = 3, Generations = 10, Population Size = 10** | **0.98** | **0.848** |
| **2** | **AutoGluon** | **Medium presets** | **0.805** | **0.78** |
| **3** | **XGBoost** | **Colsample\_bylevel = 0.908,**  **Colsample\_bytree = 0.634,**  **N\_estimators = 945,**  **eval\_metric = ‘los\_loss’,**  **Learning\_late = 0.08683,Max\_depth = 5** | **0.8902** | **0.87306** |

**AUC Graph of Best Solution**



**Notebook Submission**

The notebook with the best performance has been included in the ZIP file as **hw2\_316472455.ipynb**. It contains the complete implementation details, including data preprocessing, model training, evaluation, and SHAP analysis.

### **Additional Trials: Exploring Advanced Models**

This section outlines the additional experiments conducted using various advanced models and techniques to improve performance. These experiments focused on testing model diversity, novel architectures, and ensemble methods.

#### **1. Multi-Learner Stacking Ensemble**

##### **Methodology**

* Implemented a **stacking ensemble** with **10 base learners**, including:
  + Logistic Regression, Random Forest, LightGBM, XGBoost, CatBoost, Extra Trees, Decision Trees, and K-Nearest Neighbors.
  + Non-traditional learners like Gaussian Mixture Models (GMM).
* **Out-of-Fold (OOF) Predictions**:
  + Used **Stratified K-Fold Cross-Validation** to generate OOF predictions from each base learner to avoid overfitting in the meta-learner.
* **Meta-Learner**:
  + A Logistic Regression model trained using both OOF predictions and original features.

#### **2. 1D Convolutional Neural Network (CNN)**

##### **Methodology**

* Treated the tabular data as **1D sequences** and built a simple **1D CNN** for binary classification.
* Model architecture:
  + 3 convolutional layers with **ReLU activations** followed by **MaxPooling**.
  + Final dense layer for binary classification with **softmax** activation.
* Data split:
  + Used Tabular1DDataset to format features for PyTorch and trained using a train-validation split.
* Loss Function: Cross-Entropy.
* Optimizer: Adam.

##### **Insights**

* 1D CNN effectively captured feature interactions in a data-driven manner.
* However, the model was computationally more expensive than traditional ML models and required fine-tuning for optimal performance.

#### **3. Diversified Ensemble Models (SDEModel)**

##### **Methodology**

* Designed an ensemble model focusing on **model diversity and complementarity**.
* Steps:
  + Trained a large pool of diverse models (e.g., CatBoost, Random Forest, Logistic Regression, Naive Bayes, KNN).
  + Used **forward selection** to iteratively add models to the ensemble that maximized **AUC improvement**.
  + Assigned weights to models based on their AUC contribution.
* Subsampling:
  + Randomized row and column subsampling to improve diversity in individual model training.
  + Selected models using techniques like **Diversity, Cosine, AUC Coverage** and **K-Means Clustering** on predictions.

##### **Insights**

* The POSDModel's focus on diversity yielded a highly generalized ensemble.
* Subsampling strategies significantly enhanced model variety and reduced overfitting.

### **Overall Comparisons**

|  |  |  |
| --- | --- | --- |
| **Method** | **Validation AUC** | **Key Insights** |
| Stacking Ensemble | 0.866 | Meta-learner effectively captured diversity from OOF predictions. |
| 1D CNN | 0.84 | Showed potential but expensive to keep experimenting with. |
| POSDModel | 0.866 | Showed great potential with theshold + diversity selection technique |