Assessment 1 Individual Reflection by Yifan Wang

**Introduction**:

This project aims to develop a predictive model that uses different algorithms for binary classification to determine whether an individual is likely to drink alcohol. It involves data collection, preprocessing, model building and evaluation. The dataset we use is *Smoking and Drinking Dataset with body signal* from Kaggle. The goal will be to find a standard way to evaluate models by comparing respective works''.

I selected an appropriate model based on the characteristics of the data set and used R for fitting, including some EDA work. These parts account for about 50% of the R code. Also participated in group decision-making meetings.

**A brief review:**

1. **Data selection**. Targeting the Physical examination data has proven to be a very good choice; Our dataset contains 240,000 samples, including gender, age, height, weight, waist circumference, blood routine, vision, hearing, kidney function indicators, liver function indicators and other major physiological and health-related variables. These characteristics may be potentially related to drinking behaviour. This makes it a suitable basis for research and building predictive models.
2. **Model selection**. XGBoost (Extreme Gradient Boosting) is a widely used gradient boosted tree algorithm that builds a set of decision trees by continuously iterating over a training set, with each tree trying to correct the errors of the previous tree. Compared with traditional methods adds Gradient Boosting

$$g_i = \frac{\partial L(y_i, p_i)}{\partial p_i} = \frac{y_i - p_i}{p_i(1 - p_i)},$$

employs the gradient boosting algorithm to train the model. In each iteration, it computes the negative gradient with respect to the loss function and then fits a tree to approximate this negative gradient.

I chose this method because in theory XGBoost has higher accuracy and speed with the support of advanced algorithms. I set its objective function to "binary:logistic" . In addition, its other parameters such as binary:logitraw, binary:logitboost can also be selected, but I chose the most commonly used settings.

The "binary:logistic" objective function estimates probabilities in binary classification problems by using a logistic regression model. The model output is a probability value between 0 and 1, indicating the probability that a sample belongs to the positive category. Generally, if the probability is greater than or equal to 0.5, the sample is classified as a positive category, otherwise it is classified as a negative category.

It is worth noting that the performance of the model is not fixed, nor does it increase monotonically as the number of iterations increases, but slowly decreases after reaching a peak at about 90 rounds. The data here are taken as the best performance of the model.
3. **Result evaluation**. In our research, we unanimously consider recall as the most suitable metric for model evaluation. Generally, a high recall indicates that the model is proficient at identifying actual drinkers, thereby minimizing the risk of missing them. Alcohol consumption often correlates with concealed health issues. Accurately identifying whether a customer consumes alcohol is crucial for determining factors such as insurance costs or policy duration. My model achieves a recall rate of 73.8%, which is commendable, given the nearly balanced proportion of drinkers and non-drinkers in the dataset.

**Overall**. In my perspective, the decisions made by our team throughout the project demonstrate sound judgment. This includes our choice of datasets and the utilization of diverse models to showcase classification methods that align with the project's requirements, showcasing strong collaborative efficiency.

**Improvements and future work**

i. **Preprocessing of data sets.** Although our data set provides very complete information, I did not do enough pre-processing on it. For example, I used all features for model fitting, but this would complicate the model and increase the number of calculations. Less relevant features, such as vision and hearing, should be appropriately exclude.

ii. **The correlation between features.** Another problem is that needs improvement is that it does not consider the correlation between characteristics. For example, height and weight are often highly correlated with gender, which leads to the interesting conclusion that height is the biggest factor affecting whether to drink alcohol. It would be better to use PCA and other methods first before fitting the model.

**iii.** **Understanding of the model.** I'm not fully familiar with using XGboost, which means I can only set my parameters according to the standard situation. If I could understand this model better, I might be able to fine-tune the parameters to fit our dataset.