**Milestone P3:**

1. **Title:** Comparing Random Forest with Logistic Regression to predict a client ‘s subscription of a term deposit
2. **Abstract**

The previous paper compares the logistic regression and random forest in predicting the rare political events, civil war. Also, the authors analyze the variable importance for random forests such that to determine the causes of civil war onset. We propose to study the power of random forest in predicting rare events in another field. We will use a dataset recording direct marketing campaigns (bank term deposit) of a Portuguese banking institution. The subscriptions of the bank term deposit are based on phone calls, so this is a rare event. We plan to apply random forest and logistic regression method to predict if a client will subscribe a term deposit in the bank, and compare their predictive accuracies. Then, we will analyze the causal effects in the subscription of a term deposit from a bank. Finally, we could indicate which clients are more likely to subscribe for term deposits.

**Research questions**1. What is the accuracy of random forests in predicting the rare event in a new filed?

2. Can random forests predict the subscriptions of the bank term deposit efficiently and generally?

3. To what extent does random forest outperform logistic regression method in predicting the rare events, subscription of the bank term deposit?

4. Is there a causal effect of a client subscribing a bank term deposit?

5. What is the most important factor affecting a client subscribe a term deposit? Which clients are more likely to subscribe a term deposit?

1. **Proposed datasets**
   * bank-additional-full.csv from “Bank Marketing Data Set” UCI machine learning repository. There are 41188 observations each with 20 features such as a client’s age, sex and job. The observations are ordered by date (from May 2008 to November 2010). The dependent variable is ‘y’ indicating whether a client subscribe the bank term deposit.

The dataset is highly imbalanced because few people subscribe the bank term deposit. The ratio of subscribed (‘Yes’) bank term deposit and not ('no') subscribed in the data is roughly 1:70. In addition, there are missing or unknown values such as ‘education’, ‘housing’ and ‘loan’ in the dataset because some people would not like to tell their private information to the banking representative. We will discard the rows containing missing values.

1. **Methods**  
   **Data collection:** (a) Collect the dataset from UCI Machine learning Repository (b) Discard the rows containing missing values in the dataset (c) Determining the categorical features and numerical features (d) Visualize the relationship between features and dependent variable y (e) get the proper features
2. **Building the logistic regression model:** After getting the proper dataset and features, we will use these data to build three logistic regression models with no regularization, L1 norm and L2 norm regularization, respectively. Then we use ten-fold cross validation to plot the ROC curve, and calculate the AUC score and F1 score of each model. Finally, we will get the best model.
3. **Building the random forest model:** We will implement random forest model. After selecting proper features, we will build a random forest model. We will test different hyperparameters, and use ten-fold cross validation to calculate the average accuracy with 95% confidence intervals. Then, we select the best hyperparameters.
4. **Compare logistic regression model and random forest model:** We use ten-fold cross validation to plot the ROC curve, and calculate the AUC score and F1 score of the best logistic regression model and the best random forest model respectively.
5. **Data analysis:** We will visualize the variable importance by mean decrease in Gini Score of random forest model. Second, we analyze the importance of the features. Third, we will draw the partial dependence plots. We will analyze how the percentage of total votes for subscribing the bank term deposit changes over the range of each predictor. Finally, we summarize the causes of subscribing bank term deposits.
6. **Proposed timeline**  
   **Week 1:** Collecting data from UCI machine learning repository, and then clean them. Also, we visualize the relationship between each feature and dependent variable. Then we will try to select the proper features.

**Week 2:** Implementing logistic regression models with different hyperparameters, and select the best one. Implementing random forest models with different hyperparameters, and select the best one. Comparing the predictive performances of the best logistic regression model and random forest model.

**Week 3:** Visualize the variable importance by mean decrease in Gini Score of random forest model. Then, we will draw the partial dependence plots. Finally, analyzing the causal effects of subscribing the bank term deposits.

1. **Organization within the team**
   * Bohan will handle downloading, merging and enriching the dataset in week 1. The resulting code and datasets will be shared with the other members of the team. Data will be washed and preprocessed at that time. In week 2 (or earlier if time allows it), Zengyu will implement the logistic regression algorithm on the dataset. Qunyou will focus on using random forest method to analyze data. They will build proper model and compare the given result. The result will be recorded for further process.
   * At the end of week 2, Bohang will try to visualize these results in an intuitive way and draw graph to represent the given prediction and comparing result. Then zengyu and Qunyou will focus on writing the data story and preparing all needed figures and examples.
   * In week 3, we will work together, discuss the result and give conclusion. And we will try to make a video together and prepare a presentation in this video. We will try to attractive method to express our conclusion and give audience some useful information.