**Milestone P3:**

1. **Title:** Will client subscribe a term deposit? Comparing Random Forest with Logistic Regression in prediction
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. In the paper, we propose to study the power of random forest in predicting rare events in another field. We plan to apply random forest and logistic regression method to predict if the client will subscribe a term deposit in the bank, and compare their predictive accuracies. In addition, we would like to analyze the causal effects in the rare events. To do so, we will use a set new data set from internet and analyze it. We will use logistic regression firstly and then build random forest mode to predict the coming events.

After analyzing the dataset, we will visualize the result by using proper figure. This would give us an intuitive way to evaluate the performance. At last, we will compare the performance and give our conclusion at the end.

1. **Research questions**1. Is random forest a proper method in predicting if the client will subscribe a term deposit?  
   2. Can logistic regression give a proper solution in this field?  
   3. To what extent does random forest outperform classical method?
2. **Proposed datasets**
   * bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
   * bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
   * bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
   * bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
3. **Methods**  
   **Data collection:** XXXXX  
   **Building the linear regression model:** After getting the data, we will use these data to build a model by using linear regression. Then we use this kind of model to predict health event and compare it with real data.

**Building the random forest model:** We will implement random forest model. By selecting proper features, we will build a random forest model. Then we will use this model to predict adverse health events.

**Data analysis:** We are interested in the performance in the two above mentioned model. In this part, we will compare the result of these two prediction methods. At last we will select proper method to visualize the performance of them.

1. **Proposed timeline**  
   **Week 1:** Grabbing data from internet and then clean them before further step. Then we will try to build classical model based on it, and try to optimize it.

**Week 2:** implementing random forest and analyze the result. Then using suitable visualization method to represent the result intuitively.

**Week 3:** continuing with analysis, giving a conclusion about these two methods and preparing the data story and short video.

1. **Organization within the team**
   * 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.