Hi, today Zheyuan, Jiaqi and I will be presenting our analysis and finding on bank term deposit cold calling prediction.

So cold calling refers to the type of telemarking in which a salesperson makes an unsolicited call on a prospective customer. This method has a long history in bank marketing and has proved its effectiveness in developing the customer base. However, one huge drawback of cold calling is that it has a rather low success rate. According to statistics, only about 28% cold calls result in a conversation, and only 2% are successful. In other words, a lot of time and effort of the cold caller has gone into waste. Therefore, we were interested in if this situation could be improved with machine learning. To better guide our analysis, we raised a few questions. Firstly, can we predict if a client will subscribe to a term deposit? Also, what strategies can we use to get a client subscription? Lastly, can we build a scorecard for bank client based on their information?

The dataset we used is from Kaggle, and it contains marketing information collected by direct phone calls to evaluate whether clients subscribe to a bank term deposit for a Portuguese banking institution.

As for the variables, they can be categorized into demographics, past customer events, past direct marketing contacts, campaign information and socioeconomic factors. Though the meaning of most of the variables are straightforward, as you can just tell by their names, I would like to briefly go over the confusing ones. Among the socioeconomic factors, euribor 3 month rate is short for Euro Interbank Offered Rate, which is a [reference rate](https://www.investopedia.com/terms/r/referencerate.asp) that is constructed from the average interest rate at which [eurozone](https://www.investopedia.com/terms/e/eurozone.asp) banks offer unsecured funds to other banks. Moreover, the target variable of the dataset is y, which describes if the client subscribed the term deposit or not. It is a binary categorical variable with values yes or no.

So, in order to see what the data actually looks like, we plotted the distribution of all six categorical variables with side-by-side bar charts, and the correlation between each numeric variable with a correlation heatmap. From the bar chart, we can see one issue with our data is that it is quite imbalanced in terms of the result of subscription. The number of No’s are significantly more than the number of Yes’s. Therefore, in later analysis, we would attempt to use weights and resampling as measures to counter the data imbalance. Now if you look at the heatmap, and focus on the target variable y, we can see that it has a positive correlation with the duration of the cold call. It also has a negative correlation with pdays: the number of days that passed by after the client was last contacted; Also y has negative correlations with most of the socioeconomic factors: employment variation rate, consumer price index, Euribor 3 month rate, and employment rate.