**Project report (25 marks):**

**1. Project Goal**

The goal of this project is to build an optimal model using ABC Wireless Inc.’s historical data to predict churn rate for future customers. Using this model, the company should be able to identify customers who are likely to churn with an optimal accuracy, and can then provide suitable intervention to encourage them to stay.

**2. Overview of data, including data exploration analysis**

The dataset used in this project consists of 20 variables and 3,333 observations. There is no missing value in the dataset.

Out of the 20 variables, the first 19 variables are predictors, which are listed beloew:

1. state (categorical),
2. account\_length,
3. area\_code,
4. international\_plan (yes/no),
5. voice\_mail\_plan (yes/no),
6. number\_vmail\_messages,
7. total\_day\_minutes,
8. total\_day\_calls,
9. total\_day\_charge,
10. total\_eve\_minutes,
11. total\_eve\_calls,
12. total\_eve\_charge,
13. total\_night\_minutes,
14. total\_night\_calls,
15. total\_night\_charge,
16. total\_intl\_minutes,
17. total\_intl\_calls,
18. total\_intl\_charge
19. number\_customer\_service\_calls.

The last variable, “churn”, is the target variable which takes two values: “yes” and “no”.

4 of the 19 predictors are categorical variables: “state”, “area\_code”, “international plan”, “voice\_mail\_plan”. I’m not sure we would need normalization here if we are applying logistic regression as our model. What other data preparation do you guys think we need?

[insert R code screensot here]

**3. Details of your modeling strategy (i.e. what technique and why)**

Our team believe the most fitting modeling strategy for this project is logistic regression analysis for the following reasons:

1. The goal of this model is to estimate the probability that a customer belong to either the “no” or the “yes” category in terms of churning, which make it a classification problem.
2. Furthermore, this is a supervised machine learning model. Therefore, we ruled out unsupervised machine learning models such as k-NN/k-means and hierarchical clustering.
3. Finally, between logistic regression and decision tree, both of which are supervised clustering modes, we chose logistic regression because of the binary nature of the outcome variable.

**4. Estimation of model’s performance**

**5. Insights and conclusions**

You can include snapshots of your R code and the outputs in the report (recommended). You have to submit a single document per group in PDF format. The first page of the document should include a table with a list of the names of the group participants and a very brief summary of contribution of each team member.