# MA678 Midterm Project Proposal

#### Author: Xihao Cao

#### Date: 11/2/2021

## Personal Statement:

Big Data is one of the most popular topics in society, and almost every industry meet tons of data that contain the keys to insight into new

business opportunities. Thus, as someone who can play with data professionally, a data analyst/scientist/consultant is widely needed. My career

goal is to be a professional in the data area, and I find myself interested in communicating with others, thus working in the HR department as

a data analyst may be a good choice for me. In this project, I want to stand in the HR position to understand the factors that lead a person

to leave the current job and try to predict whether a given candidate will leave the current job.

## The data set:

I get this data set from the Kaggle database, the link is https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists. This dataset

includes information of each candidate including their demographics, gender, previous working experience, education level, type of University course

enrolled, the difference in years between previous job and current job, training hours completed, etc. The outcomes are their final decisions, a binary

value indicating whether a candidate leaves the job.

## Two questions I want to ask and answer:

1. Which are the main factors that affect the decision of a candidate?

2. Can we predict whether candidates will stay if given their information?

## Proposed Timeline of work:

I plan to finish EDA before: 11/10/2021

I plan to finish Data Processing before: 11/20/2021

I plan to finish Modeling and Validation before: 11/30/2021

I plan to finish the final report before: 12/2/2021

(the timeline may change a little depedning on the actual status)

## Some concerns:

1.The data owner said the data set is imbalanced, I think I need to further explore where the problem is.

2.There are total of 13 columns, I need to analyze and decide whether I should include all of them as predictors.

This project is about conducting Exploratory Data Analysis and establishing mixed effect model to the research of burial effects on fetal piglets decomposition. The variables we use to quantify the decomposition rate are TBS score and mass loss percentage, and we build the same model twice with each of them as outcome. We divide them into groups by their burial time, which are 1,2,3,4,5,6,12,18 months long. The fixed effects are burial depth, wrapping condition, and whether burial time exceeds 6 months. After adding the 18-month data, we find that the AIC gets smaller. For the same model, we also use the Bayesian estimation to examine the credible interval which is the range containing a particular percentage of probable values.