Project 1

In this first project you will create a framework to scope out data science projects. This framework will provide you with a guide to develop a well-articulated problem statement and analysis plan that will be robust and reproducible.

Read and evaluate the following problem statement:

Determine which free-tier customers will covert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer useage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

1. What is the outcome?

Answer: Return customer indicator (yes/no)

2. What are the predictors/covariates?

Answer: Age, gender, location, profesion, days since last log in, and activity score

3. What timeframe is this data relevent for?

Answer: Jan - Apr 2015

4. What is the hypothesis?

Answer: Dempographic and customer useage data will allow us to predict if a free-tier customer will convert to a paying customer

Let's get started with our dataset

1. Create a data dictionary

Answer:

Variable	Description	Type of Variable
Var 1 Admit	0 = not admit 1 = admit	categorical
Var 2 GRE	graduate record examination	interval
Var 3 GPA	undergrad school	interval
Var 4 Prestige	school caliber	ordinal

We would like to explore the association between X and Y

2. What is the outcome?

Answer: Outcome is Var 1 Admit (Y variable): 0 for not admit or 1 for admit

3. What are the predictors/covariates?

Answer: Predictors Var 2 GRE, Var 3 GPA, and Var 4 Prestige

4. What timeframe is this data relevent for?

Answer: The time frame of this dataset is unknown. It is likely to be cross-sectional data, but there is a possibility of longitutinal (student is not admitted one year, and is admitted later on).

4. What is the hypothesis?

Answer:

Is there a positive correlation between Admissions and a student's application (GRE, GPA, and Undergrad Prestiage).

1. Using the above information, write a well-formed problem statement.

Determine which students will be admitted, using historical data of students with a wide variety of deographic data (GRE, GPA, and undergraduate school prestige).

Problem Statement

Exploratory Analysis Plan

Using the lab from a class as a guide, create an exploratory analysis plan.

1. What are the goals of the exploratory analysis?

Answer: The goals of this exploratory analysis is to analyze the data set to summarize characteristics. The main goal is to see what the data can tell us beyond formal modeling or hypothesis testing.

2a. What are the assumptions of the distribution of data?

Answer: The assumption is that the distribution of data is normal

2b. How will determine the distribution of your data?

Answer: Histogram gives a a clear visual if our distribution is normal

3a. How might outliers impact your analysis?

Answer: Outliers may impact my analysis by adding bias or skewing my dataset.

3b. How will you test for outliers?

Answer: Other ways to tets outliers are scatterplots, box plots, and the describe function

4a. What is colinearity?

Answer: Multicollinearity occurs when one predictor variables in a multiple regression model can be linearly predicted from the others

4b. How will you test for colinearity?

Answer: We can test for colinearity through seeing if there is a correlation between my predictor variables.

5. What is your exploratory analysis plan?

Using the above information, write an exploratory analysis plan that would allow you or a colleague to reproduce your analysis 1 year from now.

Answer: 1) Look at the data set through the describe function. See what the mean, mode, median values are and see if the data is normally distributed 2) Check to see if the data set has outliers. We can test this by looking at the min/max vs. the mean, mode and median values. We can visually test this through a histogram or scatterplot. 3) Lastly, check to make sure multicollinearity does not exist. Test this by running correlation analysis among my predictor variables.

Bonus Questions:

1. Outline your analysis method for predicting your outcome
2. Write an alternative problem statement for your dataset
3. Articulate the assumptions and risks of the alternative model

In [ ]:
```
#1) To predict my outcome (admission), I will use linear regression to see if
 there is a positive correlation with admissions and GRE, GPA, and undergrad p
restige.
#2) Determine if students with High Presitge have a higher admissions rate tha
n students with Low Prestige, regardless of GPA and GRE.  Another way to put i
t, determine if Prestige is a higher predictor of admissions compared to the o
ther two variables.
#3) Assumptions: Assume that the dataset is normal and does not contain outlie
rs that potentially skew the data. Assume that the predictors are not correlat
ed (multicollinearity does not exist). Risks: The data set may not be normal a
nd have bias and skewness. There is potential for multicollinearity to exist.
```