**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 usage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

DATASET: admissions.csv

**1. What is the outcome?**

Answer:

To determine the likelihood of a free – tier customer to become a paying customer based on demographic and usage data.

**2. What are the predictors/covariates?**

Answer:

Age, gender, location, profession, days since last log in, activity score

**3. What timeframe is this data relevant for?**

Answer:

Jan-April 2015

**4. What is the hypothesis?**

Answer:

Female active users with professions in demand, over 25 years old, are more likely to convert from free-tire customer to paying customers.

**Let's get started with our dataset**

**1. Create a data dictionary**

|  |  |  |
| --- | --- | --- |
| Answer:  **Variable** | **Description** | **Type of Variable** |
| admit | Indicate whether the candidate was admitted (1) or not (0) | Binary/Categorical |
| GRE | GRE score | Integer/Categorical |
| GPA | Grade point average | Float/Continuous |
| Prestige | Rank of applicant’s undergraduate alma mater 1 highest, 2 high, 3 low, 4 lowest | Categorical |

**2. What is the outcome?**

Answer:

To create a model that can determine the likelihood of an applicant to get accepted in graduate program based on GRE, GPA and Prestige data.

**3. What are the predictors/covariates?**

Answer:

GPA, GRE, Prestige

**4. What timeframe is this data relevant for?**

Answer:

Time frame has not been specified in the dataset.

**5. What is the hypothesis?**

Answer:

The study will identify associations between applicants’ GPA, GRE score, undergraduate school’s prestige and admission to graduate school.

The hypothesis is that probability of being accepted into a graduate program is higher for students who attended a top ranked undergraduate college as opposed to a lower ranked school.

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

**Problem Statement**

Based on UCLA's Logistic admission data of 400 graduate school students we want to determine the admission result (1 for admitted and 0 for rejected) based on 3 variables (the GRE score, the Grade Point Average and the rank of an applicant's undergraduate alma mater -ranging from 1 to 4, highest to lowest respectively).

**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:

To analyse the dataset and to summarize the main characteristics

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

Answer:

Normal distribution of data

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

Answer:

By plotting data – observe skew, kurtosis

**3a. How might outliers impact your analysis?**

Answer:

Outliners affects mean and std – yields incorrect mean estimation

**3b. How will you test for outliers?**

Answer:

By plotting or using interquartile range method

**4a. What is collinearity?**

Answer:

Collinearity refers to a linear relationship between two [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable). Two variables are perfectly collinear if there is an exact linear relationship between the two, so the correlation between them is equal to 1 or −1

Collinearity means the collinear variables contain almost the same information so one can safely be dropped.

**4b. How will you test for collinearity?**

Answer:

Calculate correlation between pairs of predictors

**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. Run describe function on the dataset (get mean, median, max and std deviation)
2. Create box plots and scatter plots to look for outliners
3. Look for collinearity – relation between pairs of predictors

**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