Saturday, 24 March 2018

GA Data Science Homework Week 1 (Dean Wan)

Project 1

Part 1 ("Free-tier customers")

1. ***What is the Outcome?***
   The outcome is to determine the probability of a free-tier customer converting to a paying customer based on demographic data collected on sign up and customer usage data collected from Hooli.

2. ***What are the predictors/covariates?***
   The predictors are age, gender, location, profession, days since last log in and activity score.

3. ***What timeframe is the data relevant for?***
   The data cover the period from January 2015 to April 2015.

4. ***What is the hypothesis?***
   Free-tier customers that have a lower days since last log in and higher activity score are more likely to become paying customers.

Part 2 ("Let's get started with our dataset")

1. ***Create a data dictionary***

Variable	Description	Type of Variable
admit	0 = not admitted 1 = admitted	Categorical
gre	A number between 220 and 800 (in the dataset) that indicates performance in the GRE test (higher is better)	Continuous (assuming GRE can be measured more precisely than what's in the data)
gpa	A number between 2.26 and 4.0 (in the dataset) that indicates academic performance in subject assessment (higher is better)	Continuous (assuming GPA can be measured more precisely than what's in the data)
rank	A number between 1 and 4 that indicates the level of prestige of the student's undergraduate university (lower is more prestigious).	Categorical

2. What is the outcome?

The outcome is the probability a given student will be admitted into graduate school given the prestige of their undergraduate university and their GRE and GPA scores.

3. What are the predictors/covariates?

The predictors of admittance probability are the student's GPA score, GRE score and prestige of undergraduate university.

4. What timeframe is this data relevant for?

Unknown.

5. What is the hypothesis?

Students that have higher GRE and GPA score as well as higher undergraduate university prestige will be more likely to be admitted into graduate school.

Part 3 ("Exploratory Analysis Plan")

1. What are the goals of exploratory analysis?

The goals of exploratory analysis are to gain familiarity with the data structure and distribution, giving us an idea of what the data "looks/feels" like as well as how we might classify and work with the data we have.

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

Generally speaking, there is an assumption that large sets of data are normally distributed or are somewhat symmetrical. We also generally assume that the sample we take is representative of the population.

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

We can get a general idea of the distribution of data by using summary statistics and by visualising the data. We can take measurements like the mean, mode, median, standard deviation and interquartile range, combining this with a visual plot of the data through things like box plots, histograms and scatter plots to get an idea of how our data looks in aggregate and identify outliers.

3a. How might outliers impact your analysis?

Outliers can skew the interpretation of data, which creates misleading analysis, models and conclusions.

4a. What is collinearity?

Collinearity indicates that two predictor variables have a linear relationship between each other (i.e. they are not independent) in a regression model.

4b. How will you test for collinearity?

We can test for collinearity by performing analysis on the correlations between the predictor variables in our dataset (e.g. get an R^2 score between two variables, or use ANOVA or t-tests).

5. What is your exploratory analysis plan?

Using the Admissions dataset as an example, the exploratory analysis plan would be:

1) Examine the raw dataset using a .head() method, which will give us an idea of the available data (i.e. the data types of each column as well as what type of information is available - here, we have GRE, GPA and Ranking, which are all numerical data types).

2) Using our hypothesis as the base, examine which columns we would actually want to use or are useful for the purposes of our analysis. In the Admissions dataset, probably all the data is logically going to impact the probability of admission, so we'd include all of them in our analysis.

3) Do a check for the data quality within columns (i.e. how many rows have nulls or blanks or inconsistent formats) as well as across columns (e.g. if there are two date columns, are they formatted in the same manner?). Clean, reformat or exclude columns based on the findings. In the Admissions dataset, there's no missing data and everything looks to be in a consistent format, so we don't have to worry about this step.

4) Use summary statistics and data visualisations to get a better understanding of the distribution of data as well as identify any potential outliers we may need to exclude. We'd run this for each of the columns in the Admissions dataset, looking at things like mean, mode, median, standard deviation as well as use a box plot to determine where the majority of observations fall.

5) Run R^2 correlations against other variables to determine collinearity - i.e. there's probably a fairly strong linear relationship between GRE and GPA, and determine how to deal with this.