Capstone Project – Data Wrangling

I looked at a couple of data files – ‘train\_users’ and ‘sessions’.

To get a feel for the data, I ran the info, head, tail and describe methods. I summarized the data by using the value\_counts methods to generate the frequency counts for categorical data. I also looked at the histogram plots for these variables.

The sessions data had web history information about the users with regards to the actions that they had taken. They also had the time (number of seconds) that were spent by each user. I used the groupby function to group by the action taken and summarize the time spent in terms of sum and average to compute the total time spent and average time spent by each user for each action.

I had created plots using both matplotlib and seaborn. Seaborn generated colorful plots but when there were more than 5 to 6 categories, it was hard to read the labels on the x-axis.

To impute missing values I had initially used my business knowledge. For example to impute missing values in age, I grouped it by the levels of the target variable, calculated the mean for those levels using data that was not missing and used those values to impute missing values in age by those levels.

However after discussing this with my mentor, we tried imputing the missing values using a machine learning model. To impute missing values in a continuous variable such as age I used a linear regression model and to impute missing values in a categorical variable, I used a K-Nearest Neighbors model with number of neighbors as one. I had created functions to impute these missing values. This was something new that I had learnt.

After creating dummy variables for some categorical variables, some of the new columns names were duplicates such as facebook, google etc. To differentiate them I appended the name of the source from which they were extracted such as facebook\_signup\_method, google\_signup\_method etc.

Apart from this I had created some new features such as year, month, day and hour from a timestamp variable and year, month and day from a date variable. I had created separate functions to extract these features. I did not look at outliers.