**Data Wrangling**

The takehome\_user\_engagement file was downloaded in python and the attribute time\_stamp was converted to a timestamp and grouped by week and user\_id by frequency to create the target variable adopteduser. The modified takehome\_user\_engagement was merged with the takehome\_users database file to create the dataframe for this study. All values less than three was binned into the 0-level and all values greater or equal to three were binned into the 1-level (Figure 1A in Appendix). The dataset is imbalanced and the results can be seen in the Data Quality Report (Figure 2A). The categorical attributes creation\_source was transformed to a nominal variable and the creation\_time was converted to a timestamp to create the day, month and year variables. These values can be seen in the Data Quality Report. The unique identifying variables such as user\_id, creation\_time, name, email, last\_session\_creation\_time and org\_id were not used in the analysis.

**Results**

A total of four and five input variables were used for the Random Forest, Logistic Regression Models and Extra Trees Classifier for the balanced approach. The data was split 65:35 between the training and test dataset and the variable importance values were calculated and the best f1 score for Random Forest and Logistic Model Classifier was used to determine the best model to predict future useradoption rates for the target variable. The “gini’ index was used for the Extra Trees Classifier instead of the f1 score. The variable importance for all three models was calculated to complete this assignment. The best model selected was the Random Forest Model and the top variables selected are the month, creation\_source\_t, opted\_in\_to\_mailing\_list, enabled\_for\_market\_drip and invited\_by\_user\_id\_t. The month was created from the creation\_time variable. The accuracy value was 61% and the f1 score was 0.23 (Figure 3A). When the month variable was not included in the analysis for the Logistic Regression Model, the order of variable importance was invited\_by\_user\_id\_t, creation\_source\_t, opted\_in\_to\_mailing\_list and then enabled\_for\_mailing\_list. The accuracy score for the Logistic Model was 62% and the f1 score was 22% (Figure 4A). The Random Forest Model selected a total of 245 True Positive Values compared to only 226 for the Logistic Model and did a better job in selecting the True Positive Values for 19 adoptedusers. The Extra Tree Classifier did not have the month variable included and the top four attributes were; creation\_source\_t (86.6%), invited\_by\_user\_id\_t (7.43%), opted\_in\_to\_mailing\_list (3.14%) and enabled\_for\_marketing\_drip (2.83%). The accuracy score for this model was 0.49% and did not perform as well as the other two models.

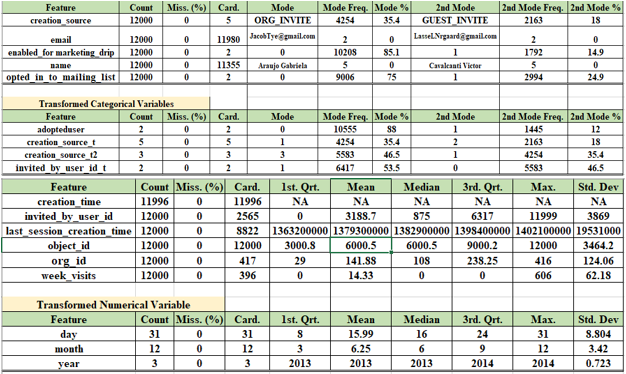
**Discussion**

The variable importance values are important in selecting only the necessary variables to create a champion model. This reduces the dimension space and can increase the metric performance used to test the model. The one area that was not investigated is the weekly distribution of use by the user\_id for this data set. The histogram was generated by the Pandas Profiler and it shows an interesting bimodal distribution between weeks 1 to 26 and then from 27 to 52 (Figure 5A). This is another area that needs to be addressed for feature selection and model improvement for this study.

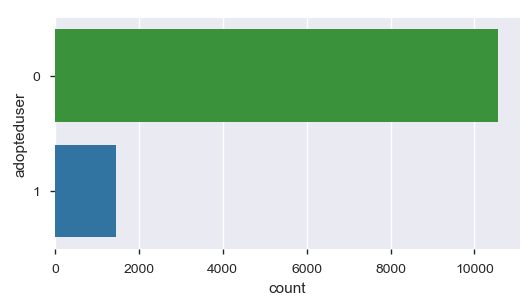
**APPENDIX**

**Figure 1A: Categorical and Continuous Data Quality Report**

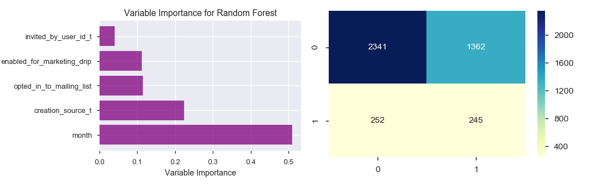
**(all\_t refers to transformed attributes).**



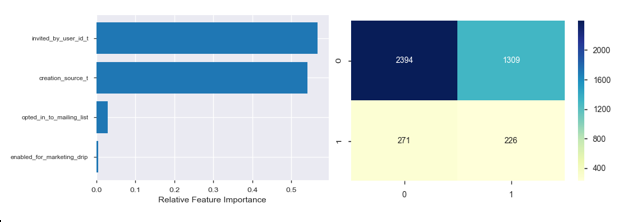
**Figure 2A: Target Variable adopteduser Count Plot.**



**Figure 3A: Random Forest Variable Importance and Confusion Matrix.**



**Figure 4A: Logistic Regression Variable Importance and Confusion Matrix.**



**Figure 5A: Weekly Distribution of user\_id and Frequency**

