**Clean data steps:**

**Excel:**

Train data:

Create minutes\_of\_day column from timestamp then put [=IFS(E2<360,1,E2<720,2,E2<980,3,E2>=980,4)] into time.

If duration\_previous = NA, then C\_previous = NA. duration\_previous = NA means there was no previous issuer. Therefore, add new variable: first\_insurance: [=IF(duration\_previous = “NA”,1,0)].

In duration\_previous: find/replace all NA with 0.

In C\_previous: find/replace all NA with 0.

Risk\_factor: 35.4% of values were NA. New variable “5” created for NA values.

**Stata:**

Use To-Do file

(repeat process for Test data then continue for only Test data)

**Excel:**

For Test data, no purchase information given, so we must keep unique customer demographics but delete the repeated demographic data for each search behavior.

Create new column called “keeper” and use value: [=IF(B1=B2,””,”keep”)].

236 entries did not have car\_value. All entries without a car value were also “NA” on duration\_previous and C\_previous, but not all “NA” on duration\_previous and C\_previous entries were blank in car\_value. Most entries with “NA” on duration\_previous and C\_previous had a car\_value of “e”, “f”, or “g” with most having “e”. The blank entries were given an “e” value.

Save as .csv and import into Stata.

**Stata:**

Command: drop if keeper != “keep”

drop keeper, customer\_id, and A-G

export excel using “C:\Users\CClifford”, firstrow(variables)

**Test the training data:**

**Python:**

Use Capstone\_Logistic\_Regression