**Aleksey Kramer**

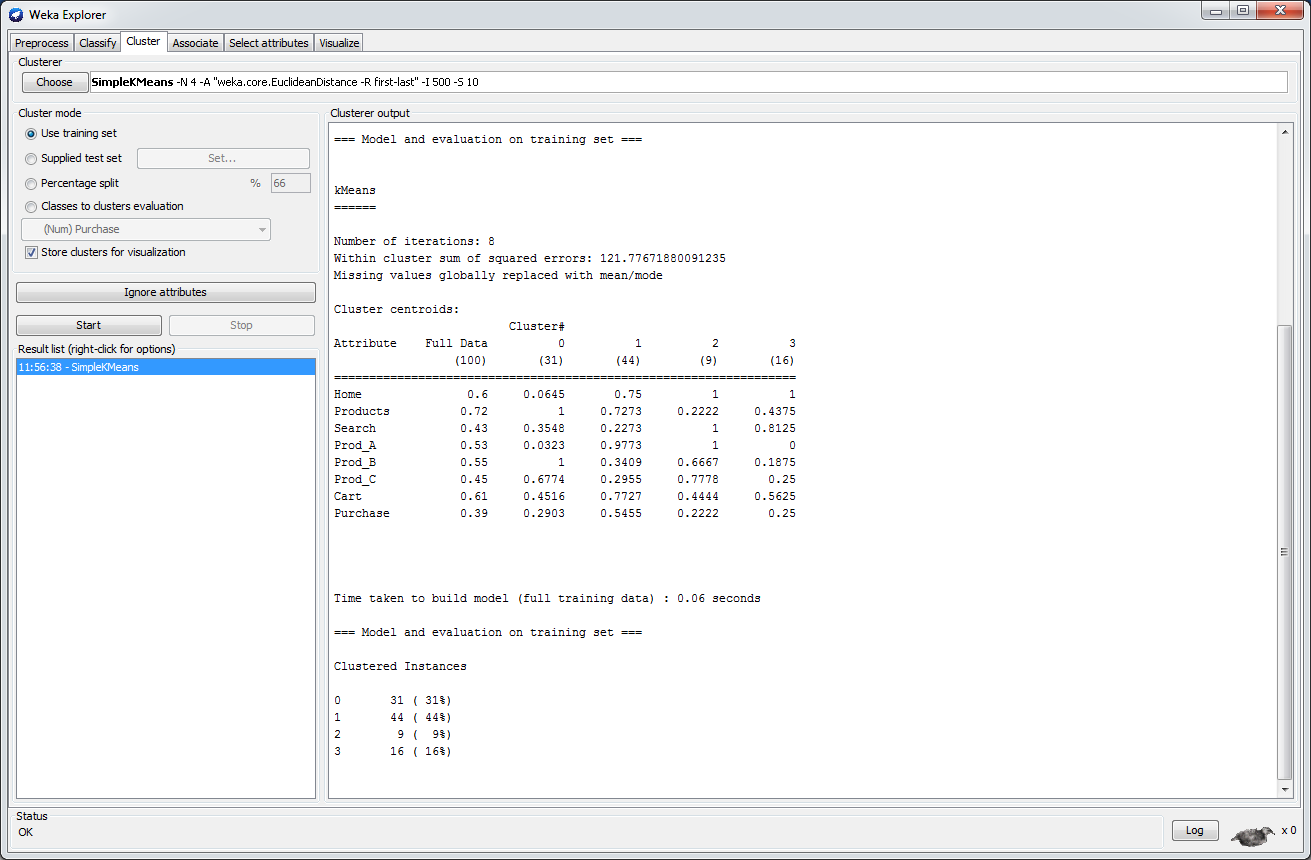
**Data Science 450**

**Homework 3 – Lecture 5**

**First part of the homework:**

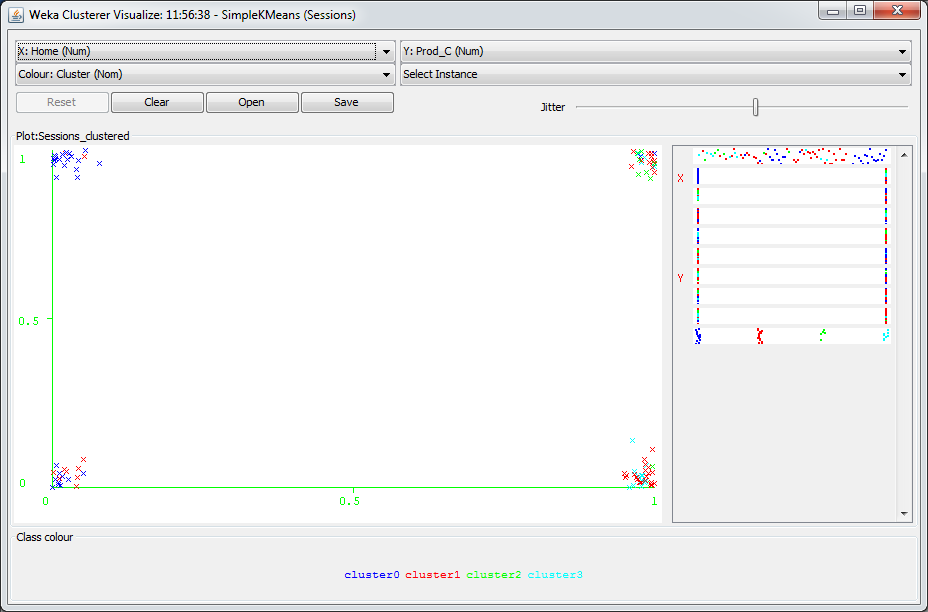
* Use WEKA's K-means clustering algorithm to cluster these user sessions into segments. Try different clustering runs with various numbers of clusters (e.g., between 4 and 8), and select the result set(s) that seem to best answer as many of the following questions as possible.

**Using 6 clusters appears to produce the best results (clusters 0 and 1 look very promising):**



* If a new user is observed to access the following pages: Home => Search => Prod\_B, according to your clusters, what other product should be recommended to this user? Explain your answer based on your clustering results. What if the new user has accessed the following sequence instead: Products => Prod\_C?

**According to the model and examining the instances of the clusters leads me to believe that Prod\_C should be recommended. Cluster 2 clearly shows that 66% of people that accessed the following path Home => Search => Prod\_B are 77% likely to buy Prod\_C. The same information can be visualized using Weka Cluster Visualization tool as shown below.**



* Can clustering help us identify casual browsers ("window shoppers"), focused browsers (those who seem to know what products they are looking for), and searchers (those using the search function to find items they want)? If so, are any of these groups show a higher or lower propensity to make a purchase?

**Clusters 1 seems to identify those who can potentially be casual shoppers. Looks like these visitors browsed the site before they figured out what they were after. (based on the cluster visualization and representation of clusters)**

* Do any of the segments show particular interest in one or more products, and if so, can we identify any special characteristics about their navigational behavior or their purchase propensity?

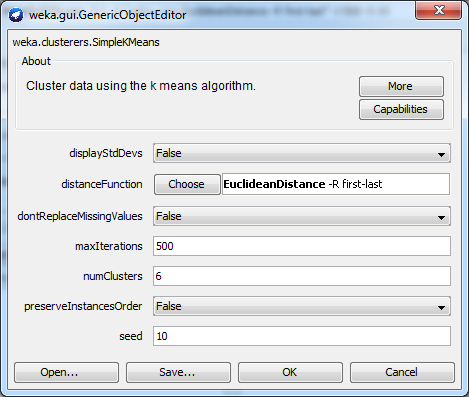
**I am not sure how to determine that. I can see that what pages shoppers visited and if the visit was followed by the purchase. For example, cluster 2 shows shoppers who visited Home -> Search -> Prod\_B , 44% added product B to the cart, but only 22% of shoppers fitting the pattern bought the product. Theoretically, the products shoppers looked at should produce a pattern of what shoppers are interested in. I am not sure how to do it in WEKA. Would it be possible to show us an example of how to do similar types of investigation in-class?**

* If we know that, during the time of data collection, independent banner ads had been placed on some popular sites pointing to products A and B, can we identify segments corresponding to visitors that respond to the ads? If so, can we determine if either of these promotional campaigns are having any success?

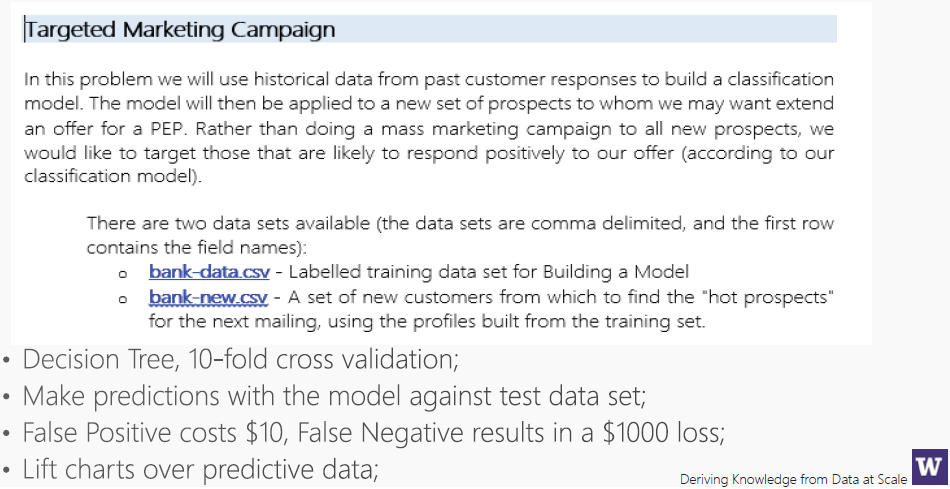
**I am not sure how to do it in WEKA, but theoretically, I would trace the paths of the shoppers to trace if there is a shortest path between home and desired product followed by the purchase during the A and B experiment time frame and compare it vs. normal traffic pattern to see which add altered shoppers’ behavior the most.**

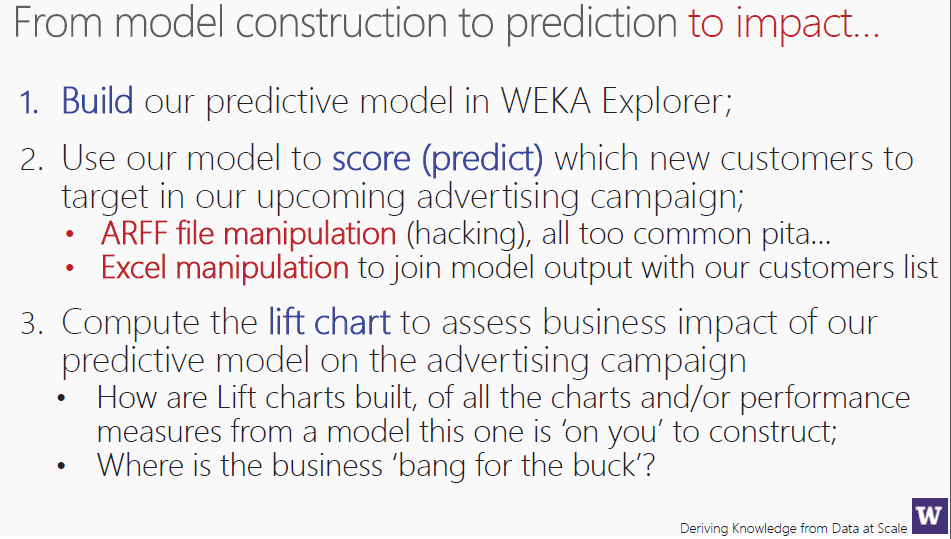
For this problem, you should submit your clustering result summary (including the cluster centroids), the final data set which shows the final assignment of these sessions to clusters, and your answers to the above questions along with your justification based on the clustering results. Other Notes: You may also want to use WEKA's cluster visualization capabilities to identify interesting distributions of various page visits among and within clusters.

**I am not sure what is meant by “… final data set which shows the final assignment of these sessions to clusters …” I am submitting the dataset and the options used to run SimpleKMeans in WEKA – it can be seen just below (6 clusters was used). Additionally, screenshots of the data exploration are presented above in the paper.**



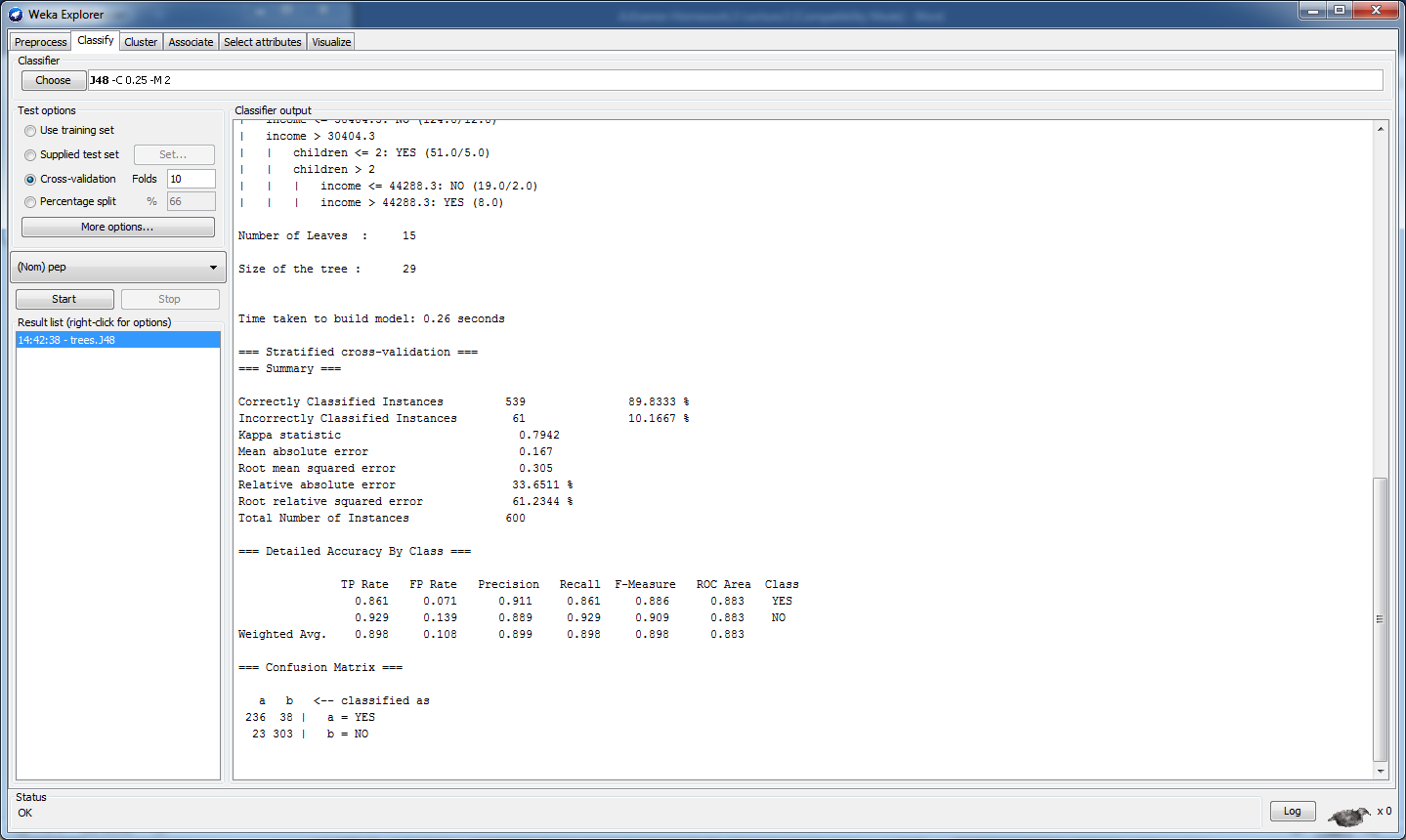
**Second part of the homework:**





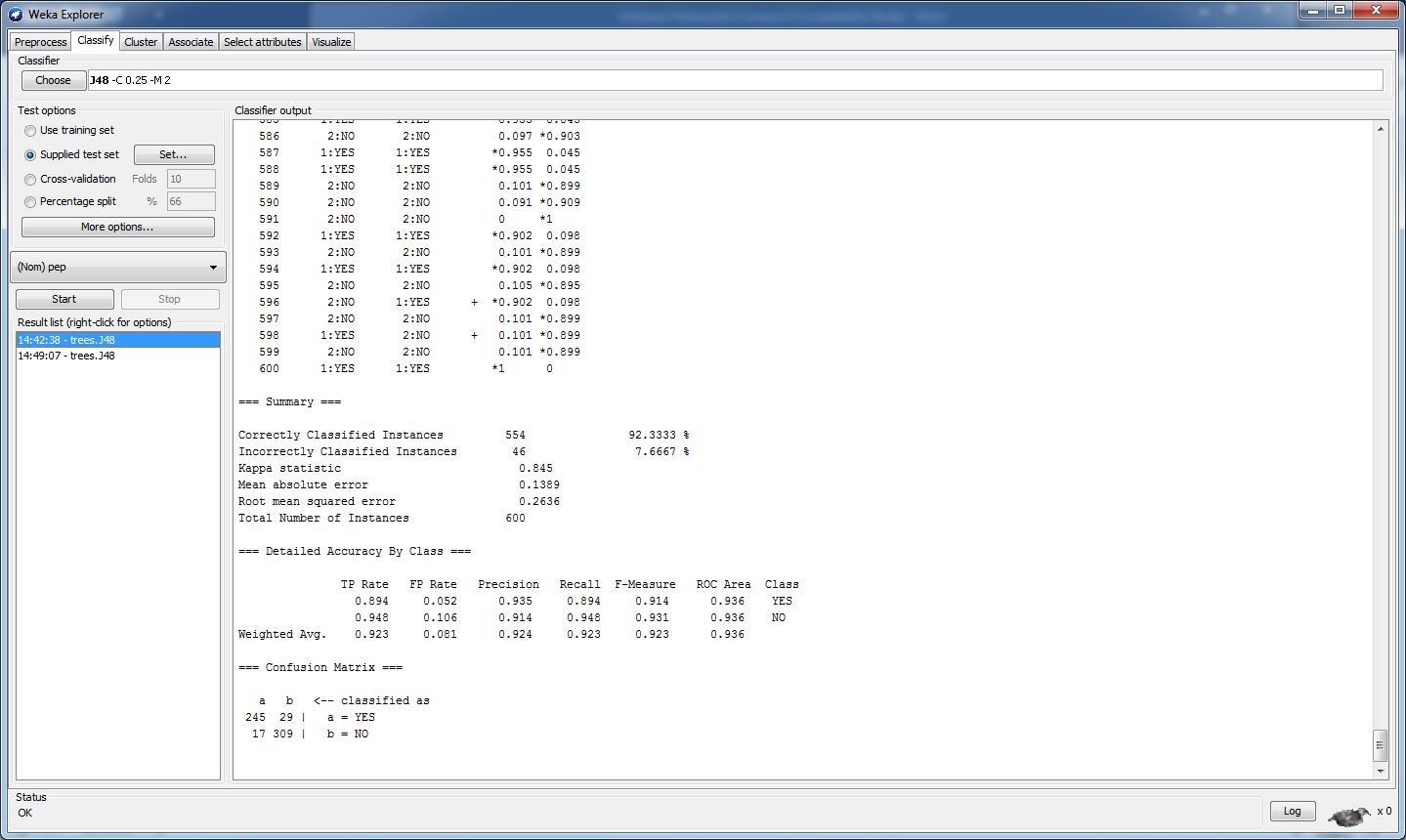
1. Create a Decision Tree using 10-fold cross validation

**Created a decision tree using 10-fold cross validation as directed in the step-by-step instructions outline above. The output – see below:**



1. Make predictions with the model against test data

**Made predictions using Bank-New data set. Actually, the test behaved much better than the model producing 92% accuracy. Actually, it is a pretty good prediction. The precision and recall seem to be good enough. See below:**



1. Lift Chart (Found in Weka - > Visualize -> Cost/Benefit Analysis -> YES). I did not have to do anything special, just read Weka’s documentation to figure out how to display a lift chart. Apparently, the Lift chart is displayed by right-clicking on the model and selecting ‘Cost/Benefit Analysis’ -> ‘Yes’ menu item. The lift char produces desired curvature. However, I am not sure how to weigh ($10 and $1000) elements in this model.

