Name _____
ID# _____

ECO 5385 Prof. Tom Fomby
Predictive Analytics for Economists Summer I, 2014

Term Project

By signing your name in the above line, you are swearing that you have received no assistance in completing this Term Project other than consulting with Professor Fomby. This Project is due at 6:00 pm on Thursday, June 26 in class. It is worth 10% of your overall grade. It should be handed in as a notebook with the pages I ask you to construct. You are to use XLMINER in completing this Term Project. If you do not have a copy of XLMINER on your laptop, you can use the computer in the lab with XLMINER loaded on it.

STEP1: Put a Cover Page in your Project Notebook with your name and title you want to choose to label the exercise, like "Used Car Prices Modeling" by (your name).

STEP 2: Download the data set "Used Car Prices_Term_Project_SumI_2014.xls" from the "Term Project" folder on my class website. This is the dataset that you will use to complete this exercise.

STEP 3: Create a second EXCEL spreadsheet in your EXCEL file by copying the columns "Price" (the target variable) and the input variables "Age_08__04", "KM", "HP", and "Weight" into the new spreadsheet. You should have 1436 cases (rows) and 5 columns. Explain the meanings of each of these variables. For more detail on this data set, see the discussion of the "Toyota Corolla" data set on pages 40 and 124 in your textbook. (See the book's subject index.)

STEP 4: Scan your data set. Are there any missing observations? If so, treat them using the "median" of the remaining observations of each variable that has one or more missing observations. You should still have 1436 cases.

Step 5: Eliminate all cases where the KM variable is **less than or equal to 20**. These observations represent brand-new cars. We only want to analyze used cars. This will of course reduce the number of cases in your previous spreadsheet and you will wind up creating a new spreadsheet that is minus the new cars.

STEP 6: Partition your "cleaned" data using a (50% training, 30% validation, and 20% test) partition of the step 5 data using the random number seed 12345. This will give you a partitioned data spreadsheet which you can use to build competing models of used car prices.

Step 7: Using the partitioned data of Step 6, build three models: K-NN, an ANN of architecture 4-2-1, and MLR using backward selection and then, among the backward selected models, the model with the maximum Adjusted R2. Use all the input variables in the K-NN and ANN models and start with the 4 inputs in the MLR procedure. Normalize the input variables in the K-NN approach. Otherwise, you don't need to normalize the input variables. Compare these models and report their respective Validation and Test Data Set RMSEs.

Step 8: Choose the best two of the above three models, thus starting the formation of a "trimmed" ensemble. Write out a description (mathematical or otherwise) of both models.

Step 9: Using the Validation data set and the two best methods that you have chosen for characterizing the used car price data (step 8), obtain the weights for a Granger-Ramanathan combination predictor. Write out the equation for the Granger-Ramanathan combination predictor.

Step 10: Using the Test data set, compare the RMSE performances of the Granger-Ramanathan ensemble, a simple average ensemble, and the two best models that you chose in Step 8. How did the ensemble models do in the hold-out (test) data set? Are the ensemble methods better than the individual prediction methods? Which is the best prediction method overall? Explain your answer.

Step 11: Retrieve the Excel file "Used Cars_Score.xls" from "Term Project" folder on the class website. Using the scoring data set that is listed in **Sheet 1** of that file, use XLMINER to score the 10 cars in that data set that are coming to auction. List the 10 scores of these cars. Hint: To score a data set using a model you have built, go to "Data Utilities" and then chose the option "Score from Stored Model." That option will ask you to select the worksheet where your stored model is and the worksheet where the data to be scored is located. This procedure will generate another worksheet with the scored prices of the cars up for bid.

Step 12: Now go to **Sheet 2** of the Excel data set "Used Cars _ Score.xls." Assume that the 10 prices listed for the 10 cars are the final bids on these cars except you have the last bid. Which car is the most undervalued? Which car is the most overvalued? If you had one car to make the last bid on, which one would it be? What would you bid for the car? Explain your answer.

Step 13: Include this Project Statement with your name signed at the top in the indicated place and include it in the back of your folder. This is your signed pledge that you have not consulted anyone other than Professor Fomby in completing this term project.