INDR 450/550 HOMEWORK 3, Due Date: April 25, 2022

- Please work in groups of two or three or (individually) and submit one file for each group with all names.
- The main data for this exercise is monthly sales of Toyota vehicles in Turkey starting in January 2006 and ending in December 2019.
- Please perform all computations in python. Please submit (upload on blackboard) your commented (with explanations) python notebooks.
- In addition to the python notebook, **submit a short typed summary report** that includes the results (error tables, prediction intervals etc.) of all exercises. Also add a general assessment of the methods (which method is the best, which should be avoided etc.). **The report is part of the overall grade.**

Exercises

Part 1: Regression and Model Reduction for Sales Prediction: The data can be found under 'ToyotaSales.csv'. We start by seperating the data into a training set (first 96 months) and a test set (months 97 to 168).

1. (35 points) Implementing a full model with all seasonal predictors and additional polynomial terms and comparing with a reduced model.

   (a) (20 points) Fit a least squares regression to training data with the following predictors:

   $$\begin{aligned} y_t &= \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 \\ &= \beta_4 x_{1t} + \beta_5 x_{2t} + ... + \beta_{15} x_{12t} + \epsilon_t \end{aligned}$$

   where x_{it} is an indicator (dummy) for month i ($i = 1, 2, , , 12$). Note that we only need to use 11 of the 12 monthly indicators in the regression. We typically skip the first month or the last month of the year but in this case, it might be better for interpretation to skip one of the months in the middle of the year (May or June) since December is a special peak month and January is an off-peak.

(b) Comment on the significance of the predictors and R^2 value. Compute the MSE, RMSE and MAPE of the fit on the training data.

(c) Use the previous model on the test data and compute the MSE, RMSE and MAPE. Comment on the issue of overfitting.

(d) (15 points) Now fit a reduced regression model that uses only one of the trend terms and only the indicators for the months of January and December:

$$y_t = \beta_0 + \beta_1 t + \beta_2 x_{1t} + \beta_3 x_{12t} + \epsilon_t$$

(e) Comment on the significance of the predictors and R^2 value. Compute the MSE, RMSE and MAPE of the fit on the training data.

(f) Compare and discuss the test error performances of the full model and the reduced model.

Part 2: Logistics Regression and Model Classificatıon for Market Directions: The data can be found under 'ToyotaSalesUp.csv'. We'll try to predict whether the market for Toyota vehicles moves up (1) or down (0) in month t (the market moves up if $y_t > y_{t-1}$, and moves down otherise). We'll use the lagged differences and dummies for months of December and January. We start by seperating the data into a training set (first 96 months) and a test set (months 97 to 168).

2. (65 points) Logistic Regression and Classification for Toyota Vehicles Market

(a) (50 points) Using the training data, fit a logistic regression to the market direction data (up or down) in month t where the predictors are Lag 1 = $y_{t-1} - y_{t-2}$, Lag 2 = $y_{t-2} - y_{t-3}$, Lag 3 = $y_{t-3} - y_{t-4}$, and Jan(indicator for January) and Dec (indicator for Dec.).

(b) Comment on the significance of the predictors.

(c) Convert the probability predictions from the model to a class prediction (1 or 0) by using a probability threshold of 1/2.

(d) Find the confusion matrix for the classification rule and comment on the error performance.

(e) Implement the logistics regression model obtained for the training set on the test set. Find the confusion matrix for the classification rule and comment on the error performance on the test set. Report the AUC measure.

(f) (15 points) Repeat the above for a reduced logistics regression model that only uses the predictors Lag1, Lag2 and Dec.

(g) Compare the AUC on the test set for the reduced model with the AUC under the full model.

Please note that the data is tricky since there are some significant ups and downs in the sales that cannot be explained with our predictors.

Part 3 (Bonus Exercise): KNN Classifier for Market Directions: We use the data from Part 2.

3. (10 points) Let's check the performance of a KNN Classifier.

(a) Implement a KNN-classifier that uses the first 96 months as the input set and the months 97 to 168 as the target set using the predictors, Lag1, Lag2, Lag3 and taking the number of neighbours $K = 5$. Find the confusion matrix and comment on the classification error rates.

(b) Experiment with the number of neighbours ($K = 1, 7, 11$) and commment on the classification error performance.