

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 10**

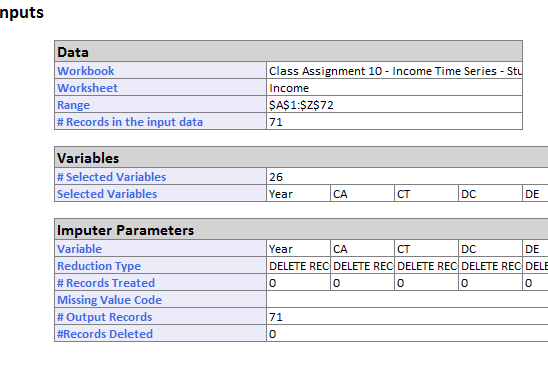
**Dr. Bharti Sharma**

**5 points**

**TASK: Time series forecasting – Data Mining Task using XLMiner**

**Perform time series forecasting on the Average Income of Tax payers in US dataset determine the best model for any of the 3 states. Provide screen shots with your explanation below.**

**Missing Data Handling Results**



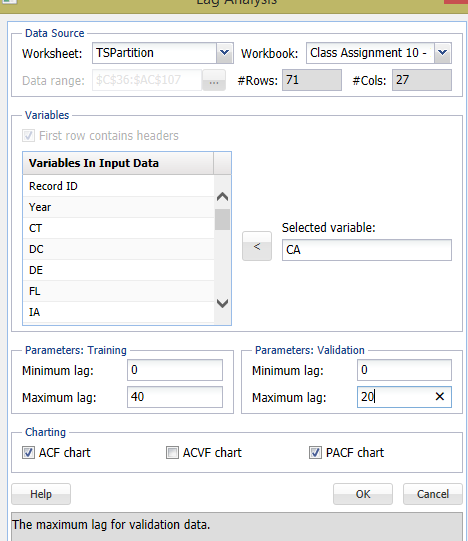
1. Perform Data Partitioning on the dataset (select Training set at least 60 % or higher).

States selected – CA, FL, NY

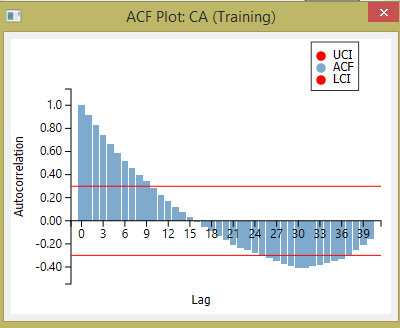
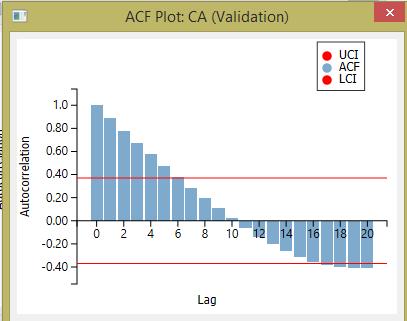
Partitions performed – 60/40, 75/25, 80/20

2. Perform Lag analysis and explain the ACF and PACF plots – you can explain in one combined comment for all 3 states. **Take Training lag of 40 and Validation lag of 20 as the max lag (lag analysis window).**

Lag analysis window



**ACF Charts (Partition: 60%/40%) – Tried do lag analysis for 80/20 partitioned data- it gave similar patterns.**

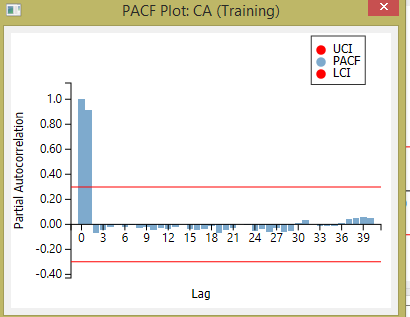
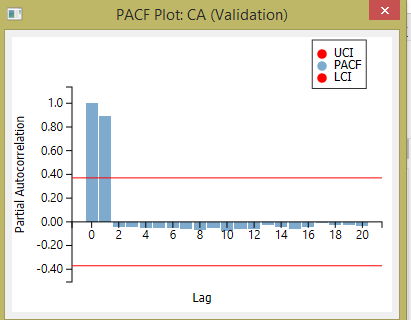
Looking at the above charts, it can be said that

1. ACF charts are similar for training and validation data

2. As autocorrelation decrease, lag increases for both, validation and training partition –means no definite pattern exists in the training or validation partition or no cyclical nature exhibited.

3. No seasonality exhibited since the pattern does not repeat in either of the two charts.

**PACF Plots**

Looking at the above charts, it can be said that

1. PACF charts are somewhat similar for training and validation data. There is a slight difference in training vs validation data after the 2nd lag, but the range of fluctuation remains from 0.05 to -0.20. Hence it can be concluded that the model is appropriate.

2. There is a pattern (first two bars(ACF1 and ACF2) are same in training and validation) meaning there is a trend that can be identified.

3. No seasonality exhibited since the pattern does not repeat in either of the two charts.

4. ACF1 and ACF2 have high partial autocorrelations. They are the same for training and validation data.

3. Develop 2 ARIMA models by changing the partitioning ‘or’ p,d,q parameters ‘or’ iterations. So total you should have 6 models (2 models for each State, choose the best model for each State).

Best model

From the below table, it can be seen that for California-Model 1 is the best,

For Florida-Model 1 is the best and for New York-Model 2 is the best. The best model has been chosen of the basis of least difference between actual versus forecast value for 1990, low validation error and low p value. The validation errors are in general high for all the models – run in millions.

|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| CA- validation error | 2,494,400.30 | 962,795,374.00 |
| CA-p value-const | 0.00 | 0.00 |
| CA-Difference actual and forecast for 1990 | 751.00 | 13,000.00 |
| FL- validation error | 4,551,364.50 | 19,121,247.00 |
| FL-p value | 0.00 | 0.05 |
| FL-Difference actual and forecast for 1990 | 1,503.00 | 2,548.00 |
| NY- validation error | 139,124,699.00 | 7,467,824.00 |
| NY-p value | 0.84 | 0.00 |
| NY-Difference actual and forecast for 1990 | 5,924.00 | 103.00 |

**Best model among all the States – New York- Model 2 because of the low difference between actual and forecasted value for 1990.**

4. Compare the models and determine the best model for each of the three states - based on the following – provide screen shots and your comments for the following model selection criteria:

a) MSE training and validation error

b) p-values of the coeff and AR1, AR2 etc

c) Forecast and Actual plot for Training

d) Forecast and Actual plot for Validation (select any year and compare the results with respect to that in all the models)

The comparisons and comments are provided in the table below. The changes made in the model are shown by the arrow

|  |  |
| --- | --- |
| **Model# 1** | **Model# 2** |
| **California CA**  **Partition: 75/25**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 200 | | **Auto-Regressive Order (p)** | 2 | | **Moving-Average Order (q)** | 3 | | **Ordinary Difference (d)** | 1 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 11537.338 |   **Validation**   |  |  | | --- | --- | | **MSE** | 2494400.3 |   **Comment: The error values are very high, which is not a good thing. The training and validation errors are far apart.**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  |  | | --- | --- | --- | --- | | **Record ID** | **Coeff** | **Std-Dev** | **p-value** | | **Const** | 10.795261 | 2.078691578 | 2.066E-07 | | **AR 1** | 0.3904963 | 0.035341782 | 2.213E-28 | | **AR 2** | 0.5635439 | 0.025519128 | 4.59E-108 |   **Comment: p values are low, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are only approximately 1000 apart, showing a better tighter fit. Hence the model is more accurate at predictions than model 2.** | **California CA**  **Partition: 75/25**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 200 | | **(p)** | 2 | | **(q)** | 2 | | **(d)** | 0 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 113297.94 |   **Validation**   |  |  | | --- | --- | | **MSE** | 962795374 |   **Comment: The error values are very high, which is not a good thing. The training and validation errors are far apart.**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  | | --- | --- | --- | | **Record ID** | **Coeff** | **p-value** | | **Const** | -0.679652 | 4.647E-16 | | **AR 1** | 2.1243312 | 0 | | **AR 2** | -1.124129 | 0 |   **Comment: p values are low, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are approximately 13000 apart. Hence not a good validation model for predictions.** |
| **FLORIDA FL**  **Partition: 75/25**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 200 | | **Auto-Regressive Order (p)** | 2 | | **Moving-Average Order (q)** | 3 | | **Ordinary Difference (d)** | 1 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 8459.0113 |   **Validation**   |  |  | | --- | --- | | **MSE** | 4551364.5 |   **Comment: The error values are very high, which is not a good thing.**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  |  | | --- | --- | --- | --- | | **Record ID** | **Coeff** | **Std-Dev** | **p-value** | | **Const** | 0.6425513 | 0.12486537 | 2.662E-07 | | **AR 1** | 1.96181 | 0.060934814 | 2.06E-227 | | **AR 2** | -0.964939 | 0.058381886 | 2.304E-61 |   **Comment: p values are low, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are only approximately 1500 apart, showing a good tighter fit. Hence the model is more accurate than model 2 at predictions**  **New York NY**  **Partition: 75/25**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 175 | | **Auto-Regressive Order (p)** | 2 | | **Moving-Average Order (q)** | 1 | | **Ordinary Difference (d)** | 3 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 5356.0669 |   **Validation**   |  |  | | --- | --- | | **MSE** | 139124699 |   **Comment: The error values are very high, which is not a good thing.**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  |  | | --- | --- | --- | --- | | **Record ID** | **Coeff** | **Std-Dev** | **p-value** | | **Const** | 2.8318025 | 14.03628354 | 0.8401132 | | **AR 1** | -0.039844 | 0.000419964 | 0 | | **AR 2** | 0.0138283 | 0.000175147 | 0 |   **Comment: p values are low or 0, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are only approximately 6000 apart, showing a good tighter fit. Hence the model is more accurate at predictions** | **FLORIDA FL**  **Partition: 75/25**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 250 | | **Auto-Regressive Order (p)** | 2 | | **Moving-Average Order (q)** | 3 | | **Ordinary Difference (d)** | 2 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 6628.8236 |   **Validation**   |  |  | | --- | --- | | **MSE** | 19121247 |   **Comment: The error values are very high, which is not a good thing. No change in values from model 1**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  |  | | --- | --- | --- | --- | | **Record ID** | **Coeff** | **Std-Dev** | **p-value** | | **Const** | 64.161829 | 32.32848021 | 0.0471796 | | **AR 1** | -0.896036 | 0.064404855 | 5.315E-44 | | **AR 2** | -0.844545 | 0.091399751 | 2.462E-20 |   **Comment: p values are low, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are only approximately 2500 apart, showing a reasonable fit. Hence the model is somewhat accurate at predictions**  **New York NY**  **Partition: 80/20**   |  |  | | --- | --- | | **ARIMA Parameters** | | | **Max #Iterations** | 175 | | **Auto-Regressive Order (p)** | 2 | | **Moving-Average Order (q)** | 3 | | **Ordinary Difference (d)** | 1 |   **a) MSE training and validation error**  **Training**   |  |  | | --- | --- | | **MSE** | 21388.516 |   **Validation**   |  |  | | --- | --- | | **MSE** | 7467824 |   **Comment: The error values are very high, which is not a good thing.**  **b) p-values of the coeff and AR1, AR2 etc**   |  |  |  |  | | --- | --- | --- | --- | | **Record ID** | **Coeff** | **Std-Dev** | **p-value** | | **Const** | -20.35376 | 3.612980726 | 1.766E-08 | | **AR 1** | 0.631919 | 0.021878517 | 1.95E-183 | | **AR 2** | 0.4412302 | 0.000188297 | 0 |   **Comment: p values are low or 0, which is a good thing.**  **c) Forecast and Actual plot for Training**    **Comment: There is a tight fitting between actual vs fitted data. This means we have good set of training data.**  **d) Forecast and Actual plot for Validation**    **Comment: Selected 1990 data for comparison. Shows they are only approximately 100 apart, showing a very good tight fit. Hence the model is most accurate at predictions than all other models.** |