Time Series Data Mining Final Project

Smaïl Darbane-Franko Sikic

April 2020

Contents

1 Introduction

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Some examples of time series are heights of ocean tides or daily temperature values. Time series are frequently used in many different fields such as: statistics, signal processing, pattern recognition, etc. and largely in any domain of applied science and engineering which involves temporal measurements.

Nowadays, high volume of data is being generated every day. Time series data mining can produce valuable information for business decisions. Possible ways to take advantage of time series dataset are:
- trend analysis
- outlier/anomaly detection
- association analysis
- forecasting
- predictive analytics

In this paper, we will approach to the airline passenger data and try to forecast the evolution of quantity of passengers. After the description of the domain in section 2, we will describe the objectives of data mining task in section 3. Furthermore, we will describe the applied techniques in section 4 and present the results in section 5. Finally, we conclude in the section 6.

2 Domain description

Current situation in the world is that everything stopped because of the Coronavirus pandemic. This situation has and will have a huge impact on companies all over the world and the economy in general. One of the industries which is the most affected by this crisis is the flying industry. Since it is a hot topic these days, we have decided to work on the data from this industry in our project.

This particular industry is worth billions of dollars and takes care of millions of passengers during a year. The development of the industry started in the first half of the 20th century and it has been growing ever since. For our project we used a AirPassengers.csv public dataset from Kaggle. This dataset provides monthly totals of a US airline passengers from 1949 to 1960. It contains of only two columns labeling the month and the amount of passengers (unit used is 'thousands'). There are 144 rows (monthly values) and the values of passengers go from 104 and 622.

Month	Thousands of Passengers
1949-01-01	112
1949-02-01	118
1949-03-01	132
1949-04-01	129
1949-05-01	121
...	...
1960-08-01	606
1960-09-01	508
1960-10-01	461
1960-11-01	390
1960-12-01	432

144 rows × 1 columns

Figure 1: AirPassengers dataset

3 Objectives of data mining task

The goal of any data mining effort can be divided in one of the following two types:
- Using data mining to generate descriptive models to solve problems.
- Using data mining to generate predictive models to solve problems.

The descriptive data mining tasks characterize the general properties of the data in the database, while predictive data mining tasks perform inference of the current data in order to make prediction. Descriptive data mining focus on finding patterns describing the data that can be interpreted by humans, and produces new, nontrivial information based on the available data set.

Predictive data mining involves using some variables or fields in the data set to predict unknown or future values of other variables of interest, and produces the model of the system described by the given data set. The goal of predictive data mining is to produce a model that can be used to perform tasks such as classification, prediction or estimation, while the goal of descriptive data mining is to gain an understanding of the analysed system by uncovering patterns and relationships in large data sets.

The goal of a descriptive data mining model is therefore to discover patterns in the data and to understand the relationships between attributes represented by the data.

3.1 Prediction and Forecasting using Time Series:

Prediction can be viewed as the construction and use of a model to assess the class of a unlabeled sample, or to assess the value or value for range of an attribute that a given sample is likely to have. According to, any of the techniques used for classification can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.
In our course we have studied several ways to make predictions based on the study of time series, how to manipulate them in order to extract as much information as possible by being able to predict the trend of our time series using statistical models, studying processes that simplify the manipulation such as the Fourier transform or other dimensional reduction processes.

3.2 Objectives of data mining Task in our project

In view of our dataset which is not voluminous, we will not have any problem at the level of the reduction of dimensionality. We do not have to apply processes as we had the occasion to use in our course such as the Fourier transform in order to approximate our time series.

But in order to have a much more accurate prediction we will try to use a process often used to know how to do transformations on our time series in order to transform into a stationary series to make the prediction much more accurate.

We will also use a statistical model which is the ARIMA model to predict the future number of passengers in this airline in the US over the next 10 years between 1960 and 1970.
Our motivation to have chosen this old dataset is to be able to compare our prediction with the reality in order to deduce if our prediction was accurate or not.

4 Applied techniques

4.1 Stationarity of time series

In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the way it changes does not itself change over time. The algebraic equivalent is thus a linear function,

perhaps, and not a constant one; the value of a linear function changes as X grows, but the way it changes remains constant — it has a constant slope; one value that captures that rate of change.

Stationarity requires the shift-invariance (in time) of the finite-dimensional distributions of a stochastic process. This means that the distribution of a finite sub-sequence of random variables of the stochastic process remains the same as we shift it along the time index axis. For example, all i.i.d. stochastic processes are stationary. Formally, the discrete stochastic process$X = \{x_i; i \in \mathbb{Z}\}$ is stationary if:

$$F_X\left(x_{t_{1+\tau}}, \ldots, x_{t_{n+\tau}}\right) = F_X\left(x_{t_1}, \ldots, x_{t_n}\right) \tag{1}$$

As mentioned earlier, before we can build a model, we need to make sure that the time series is stationary. There are two main ways to determine if a given time series is stationary:
Rolling statistics: Plot the moving mean and the moving standard deviation. The time series is stationary if it remains constant over time.
Augmented Dickey-Fuller Test (ADF): The time series is considered stationary if the p-value is low (under the null hypothesis) and if the critical values at confidence intervals of 1 per cent, 5 per cent, 10 per cent are as close as possible to the ADF statistics (Augmented Dickey-Fuller).

We will apply the first method on our time series to make sure that our time series is stationary or not.

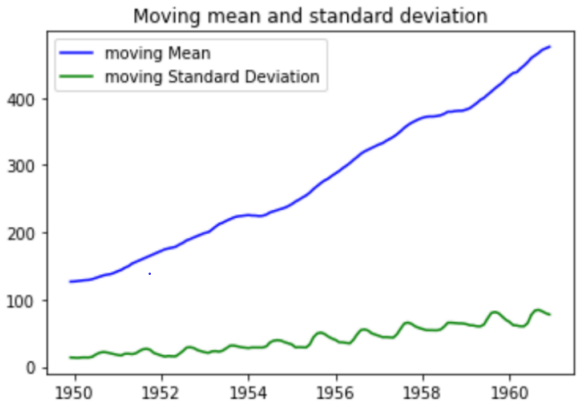

Figure 2: Moving mean and standard deviation of our TS

We can notice that the time series is not stationary because statistical averages

fluctuate over time, so we have to find a way to make our series stationary.

There are multiple transformations that we can apply to a time series to make it stationary. For example, subtract the moving average.after subtraction of the mean, the moving average and standard deviation are approximately horizontal.

The application of exponential decay is another way of transforming a time se-

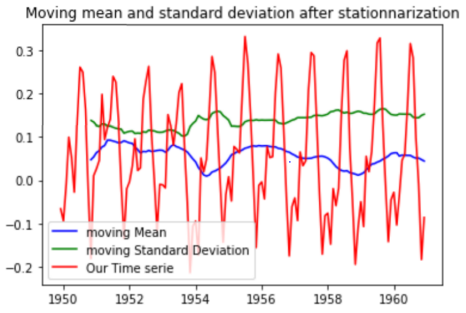

Figure 3: Application of the exponential decay to our TS

ries so that it is stationary. as we can see here method gives us a stationarity of the time series that we will have to study and therefore a much higher precision.

Linear Regression

We will compare our prediction with the ARIMA model with a classical linear regression. So before we start talking about the techniques we applied to the ARIMA model we will talk about linear regression.

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

4.2 Forecasting Model: ARIMA

What is an ARIMA model ?

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models. An ARIMA model is characterized by 3 terms: p, d, q where, p is the order of the AR term q is the order of the MA term d is the number of differencing required to make the time series stationary.

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. More on that once we finish ARIMA.
The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.
'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_1 \tag{2}$$

where, Y_{t-1} is the lag1 of the series, $\beta 1$ is the coefficient of lag1 that the model estimates and α is the intercept term, also estimated by the model.

Likewise a pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \tag{3}$$

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \tag{4}$$

To come back to our TS, we can create and adjust an ARIMA model with an AR of order 2, a difference of order 1 and an MA of order 2 :

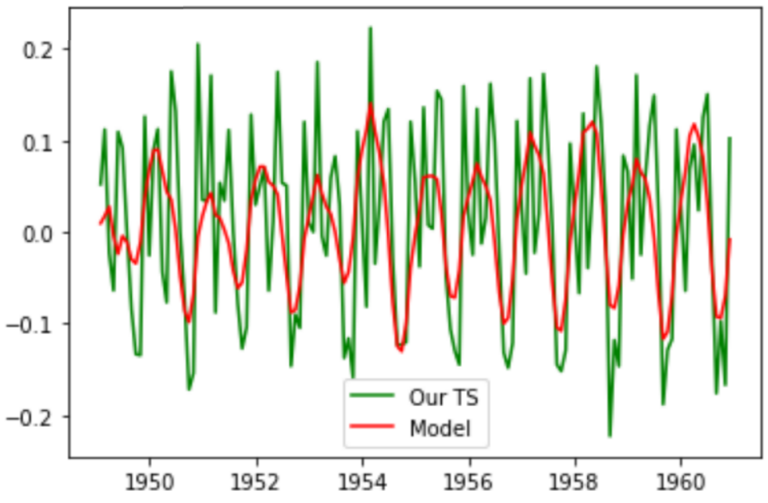

Figure 4: Comparison between the time series and our Model

5 Results

First, we used regression model in order to obtain some basic prediction results. The results of the regression are shown on figure 5. Horizontal axis represents month step from the beginning of time in the data (January, 1949). Blue dots represent values from the dataset. We used polynomial regression with degree values of 1, 3 and 5. Polynomial regression with degree value of 1 is the simplest linear regression model which produced a straight line on the graph. Model with degree value of 3 is slightly more complexed, but it produced a line which is very close to the linear regression model. Finally, the model of degree value 5 shows exponential growth.

As we can see from the graph, a 10 year prediction (so for the end of 1970) provides similar values of around 700.000 passengers for models with degrees 1 and 3, and 3.400.000 passengers for model of degree 5.

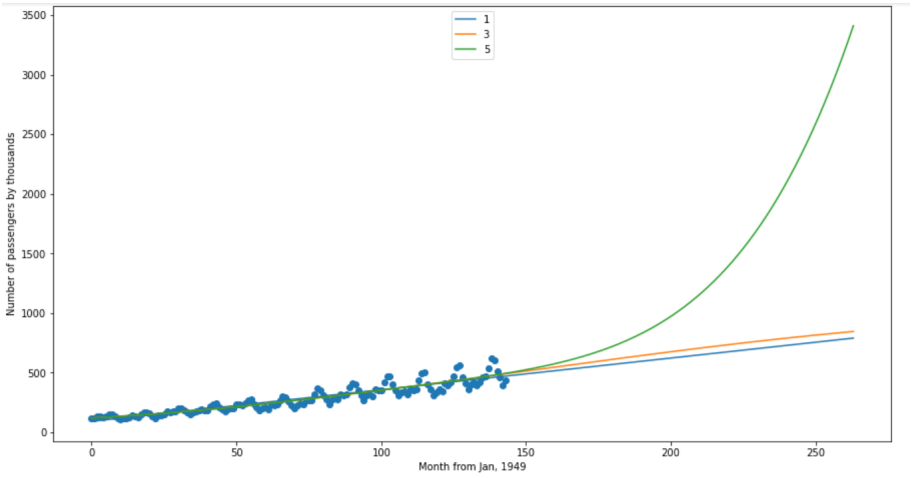

Figure 5: Forecasting using linear regression method

The results of the ARIMA model are shown on figure 6. Vertical axis represent log values of thousands of passengers. We can see that it can predict the number of airline passengers in the US over the next ten years with a 95 percent confidence zone. Specifically, it predicts that the number of passengers at the end of 1970 will be around 1.635.000, but the 95% interval goes from 735.000 to 3.295.000.
As we could see from the results, regression models provided values which simply follow either linear (or very close to linear) growth or noticeable exponential growth. The truth is somewhere in between those values and that is exactly what ARIMA model provided. But, it is interesting to notice that values provided by the regression models with degree value of 3 and 5 are really close to the 95% border interval of ARIMA model.

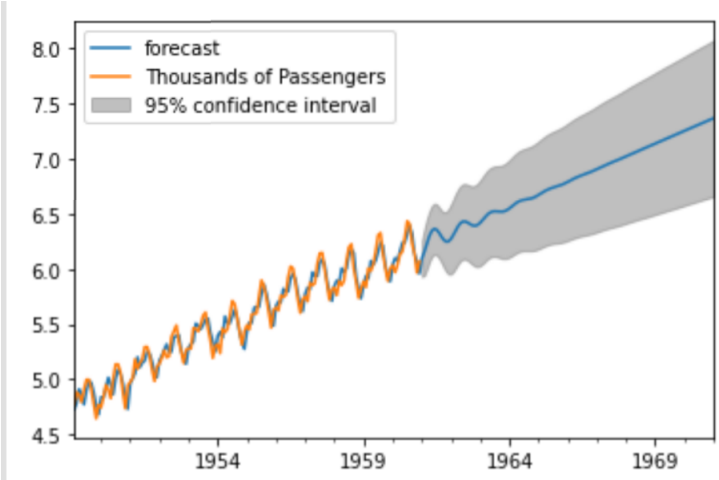

Figure 6: Forecasting the number of the airline passengers

We tried to find exact monthly data for the sixties (which was our forecasting period), but unfortunately we have not been able to find it and compare it to the result which our models provided.

6 Conclusion

In the field of Machine Learning, there is a collection of techniques for manipulating and interpreting time-dependent variables. Among these, ARIMA can remove the trend component in order to accurately predict future values. We can therefore say that in this case the ARIMA model was more efficient than the basic linear regression model. But, regression results were not that bad as we have expected.
The tools provided by the time series data mining techniques have been important to be able to master them and to achieve a more accurate prediction taking into account bias and other problems related to time series such as noise or others.

7 References

Time Series, Wikipedia, https://en.wikipedia.org/wiki/Time_series, 10.04.2020.

Shodhganga, DATA MINING TASKS https://shodhganga.inflibnet.ac.in/bitstream/, 10.04.2020.