MSDS 604: Final Project

Paul Kim, Jacques Sham, Brian Wright, and Xi Yang

December 2018

1 Introduction

Personal bankruptcy is one of the signals of economic slowdown and thus, high personal bankruptcy rate is of concern to the economy of a nation. We are interested in investigating and forecasting the personal bankruptcy rate in Canada. Our goal is to find the optimal model among different approaches to forecast the future bankruptcy of Canada.

1.1 Goal

Our goal in this report is to use different time series approaches to learn an optimal model from the training data and use this model with the test data to forecast the national bankruptcy rate in Canada between January 2015 and December 2017.

1.2 Challenge

As our data set is a time series, each data point is correlated to other data points. This correlation originates from trend and seasonality, which means that approaches such as linear regression will not work due to model assumption violations. Time series analysis will require more advanced statistical techniques to learn the models.

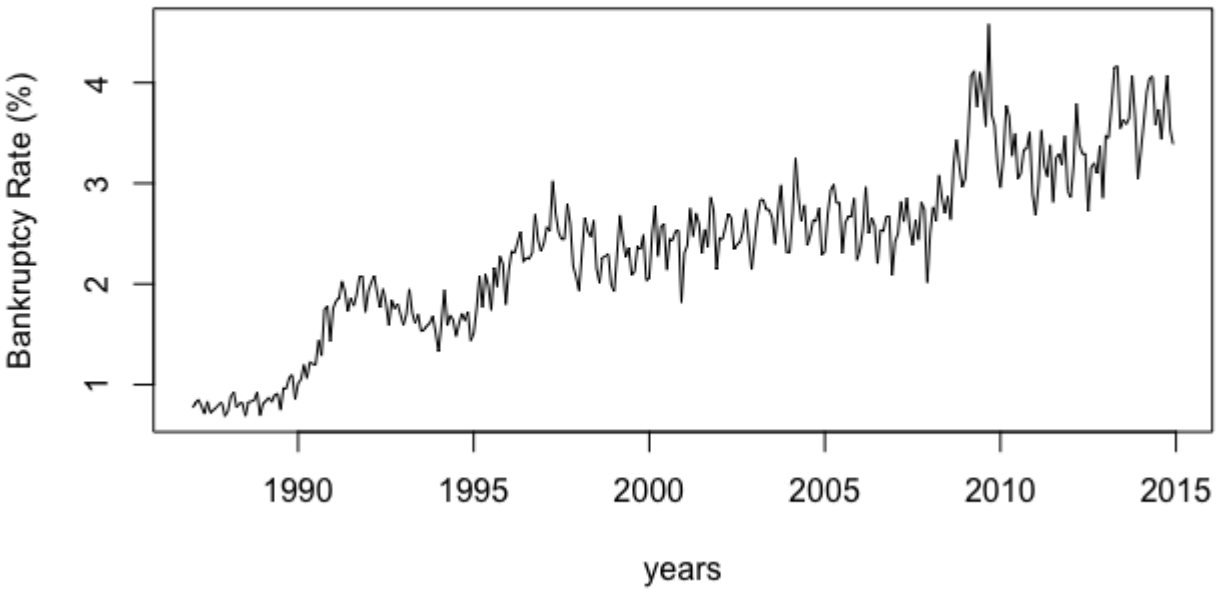

Figure 1: Canadian Bankruptcy Rate from 1987 to 2014

1.3 Terminology

There are several different terms the reader should know before reading any further in this paper. They will be defined below or upon first use of the term.

Time Series: A collection of data points corresponding to any measure of time.

Response Variable: Dependent variable (Output) of a mathematical function. In our case, the bankruptcy rate.

Explanatory Variable: Variables (Input) of a mathematical function that is believed to explain the response variable. In our case, housing-price index, population, and unemployment rate.

Trend: Consistent and general directional movement in time series describing what on average is happening.

Seasonality: Regular and predictable fluctuations based on a certain period of time.

Lag: Used to describe a data point's relationship to another in terms of time. E.g. a data point observed at time t=3 can be described as the data point observed at time t=1 with lag 2.

Differencing: The act of computing the difference between consecutive iterations of a time series.

Stationarity: A time series is said to be **stationary** if it does not possess strong trend or seasonality and therefore is easier to predict.

Training Set: The data set, time series in our report, used to learn models. In our report, the training set is from 1987 to 2010, observed monthly.

Validation Set: The data set used to measure the accuracy for each model we learn. In our report, the validation set is from 2011 to 2014, observed monthly.

Test Set: The data set of covariate information for making predictions in a given period, 2015-2017 in our report.

Box-Cox: A transformation of data that stabilizes variance. This sort of transformation is used to avoid bias in your models due to unstable variance in your data.

1.4 Approach

Bankruptcy rates in general propose a challenging problem to the world of forecasting. Forecasting bankruptcy rates can be tricky because they can have both trend and seasonality. We will explore two main types of models: univariate and multivariate models. We will use the training set data to explore each type of model with two approaches, resulting in four models. We will choose one model with the best predictive performance as the optimal model. For the purposes of prediction, we will calculate root mean squared error from the validation data and each model's predicted values for each learned model to determine predictability of each learned model.

2 Exploratory Data Analysis (EDA)

The first task of this project is to explore the data set we have. Since our goal is to predict bankruptcy rate in Canada, naturally bankruptcy rate is the response variable. Thus, unemployment rate, population, and Housing Price Index are the explanatory variables. We want to explore whether the explanatory variables are correlated to bankruptcy rate. It should be noted that the explanatory variables mentioned will not be accounted for in any univariate models.

2.1 Response Variable: Bankruptcy Rate

Figure 1 plots how the bankruptcy rate changed from 1987 to 2014. As you can see, there is a positive trend as well as seasonality in this time series, both of which need to be considered into our models.

2.2 Correlation between Explanatory Variables and Bankruptcy Rate

Two of our four model selections take account of the correlation between explanatory variables and response variable. Therefore, we want to confirm whether any relationship exists between bankruptcy rate and population, unemployment rate, and housing price index. We plot the Cross Correlation Function (CCF) between bankruptcy rate and the three explanatory variables in Figure 2. If there are spikes at lag k that significantly exceed the blue dashed line, the two variables are correlated to each other at that lag. From Figure 2, we can conclude that all three explanatory variables - unemployment rate, population and house pricing index - are correlated to the response variable. We will take this relationship into account when we learn the models.

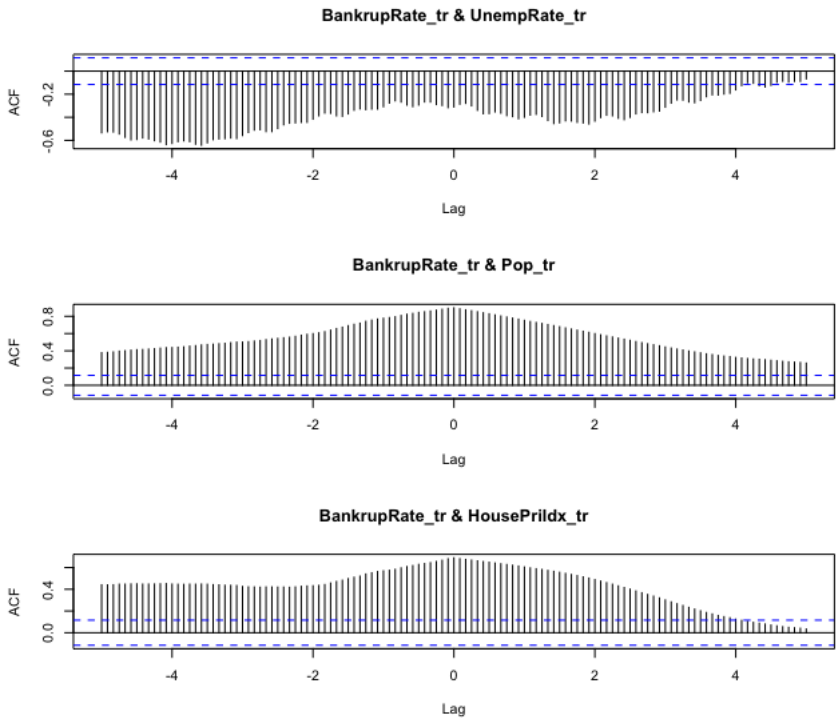

Figure 2: CCF (cross correlation function) plots

3 Available Methods

There are several different forecasting approaches that our group decided to employ. To begin, we decided to make baseline models following a univariate approach with the Box-Jenkins and Holt-Winters methodologies. These modeling approaches are described in more detail below.

3.1 Univariate Models

In regards to modeling a time series, a **univariate model** is a model with only one variable, itself. These models do not use any outside influences and are only concerned with the history of the response variable. There are two classical univariate model methodologies, the Box-Jenkins Method and the Holt-Winters Method, that will be used in this paper.

3.1.1 Box-Jenkins Method

The Box-Jenkins Method suggests learning the time series with a $SARIMA(p,d,q)(P,D,Q)m$ model, a Seasonal Autoregressive Integrated Moving Average model. This model requires fitting two ARMA models, one for the within season interactions, and one for the between seasons interactions. The p,d, and q refer to the within season model, whereas the P, D, and Q refer to the between season model. The m refers to the period of seasonality. Respectively, the p/P refer to the degree to which the Autoregressive component is present in the time series, and the q/Q refer to the degree of Moving Average needed in the model. Finally, the d/D refer to the number of differences required to make the time series stationary. The steps followed in this methodology are as follows:

1. Check for non constant variance and transform as necessary (Box-Cox or Natural Log transformation)

3.1.2 Holt-Winters Method

The Holt-Winters Method's objective is to predict the future response with the history of observations using exponential smoothing techniques. It is a pure data-driven method, and therefore we do not need to assert any assumptions to validate this model. This model is learned by taking an exponentially weighted moving average and making predictions. The prediction is purely based on historical performance and no explanatory factors are involved.

There are three types of approaches to the Holt-Winter Method to accord for the presence of trend and seasonality:

1. Single Exponential is used when trend and seasonality both are not present
2. Double Exponential is used when trend is present but not seasonality
3. Triple Exponential is used when seasonality is present, whether trend is present or not

2. Check for seasonality and/or trend and difference as necessary (d is the number of times data must be differenced)
3. Identify p,q, P, and Q from the ACF and PACF plots of the differenced data
4. Fit the proposed model and iterate to an optimal one
5. Check residuals to verify any model assumptions
6. Forecast into the future

Since we have confirmed that trend and seasonality are present in the time series, we will use the approach of Triple Exponential Smoothing for this model.

3.2 Multivariate Models

While these models performed well, their biggest flaw was that they ignored all other variables, even though external variables may have influence on our response variable. For this reason, we decided to employ **multivariate models**, i.e. models that account for external, explanatory variables (either exogenous or endogenous) that may influence the response variable. The multivariate modeling approaches considered are called SARIMAX and Vector Auto-regression (VAR), and will be described below.

3.2.1 Exogenous: Box-Jenkins Method

Exogenous variables are explanatory variables that influence the response variable but the response variable has no influence on them. If exogenous variables are present, one can use the SARIMAX model for prediction. SARIMAX's foundation is SARIMA, previously mentioned in the univariate models section. On top of SARIMA, we will add the exogenous explanatory variables. After we fit the model with SARIMAX, we still have to perform diagnostic tests for model assumptions like we do in the SARIMA approach.

3.2.2 Endogenous: Vector Autoregressive Model

Endogenous variables are explanatory variables that influence the response variable and the response variable has influence on them. In this case, we will use the vector autoregressive model (VAR). VAR treats all variables included symmetrically and considers historical data for each variable of lag p when forecasting. p is considered the 'order' of the VAR model.

4 Model Selection

In order to select the best model to forecast future bankruptcy rate, we fit all four types of models on the training set (from 1987 to 2010) and predicted bankruptcy rate on the validation set (from 2011 to 2014). We selected our models based on the model performance and we also tested whether the model assumptions are reasonable.

4.1 Model Performance

To evaluate each model's performance, we have two types of metrics. The first is goodness of fit, based on the performance on the training set. The other is predictive accuracy, based on the predictive performance on the validation data set. For the goodness of fit, we mainly compared the variance (σ^2). Smaller σ^2 corresponds to less information loss and, in turn, a better model. Note that σ^2 does not apply to every type of model. Since our goal is to forecast the bankruptcy rate, the predictive accuracy is the most critical element for model selection. The root-mean-square error (RMSE) was applied in our report. RMSE is frequently used to measure the difference between values predicted by a model or an estimator and the values observed. The smaller the RMSE is, the better the prediction is.

For each type of model, we fit multiple models and selected the best one primarily based on the RMSE and σ^2. The metrics of the best model for each type are reported in Table 1. From Table 1, the SARIMAX(4, 1, 5)*(2, 1, 2) achieved the lowest RMSE. We consider this as our best model. Figures 3 and 4 show how this model performs on the training set and validation set. The red line indicates fitted values based on our model and the black line indicates observed values. We can see that the red line is very close to the black line. Our model performs well on the training set. In Figure 4, the blue lines represent the predicted values and red line represents the observed values on validation set. Also, the blue line is very close to the red line. The shadows

indicate the prediction intervals. To further evaluate our model, we need to verify whether it fits the model assumptions.

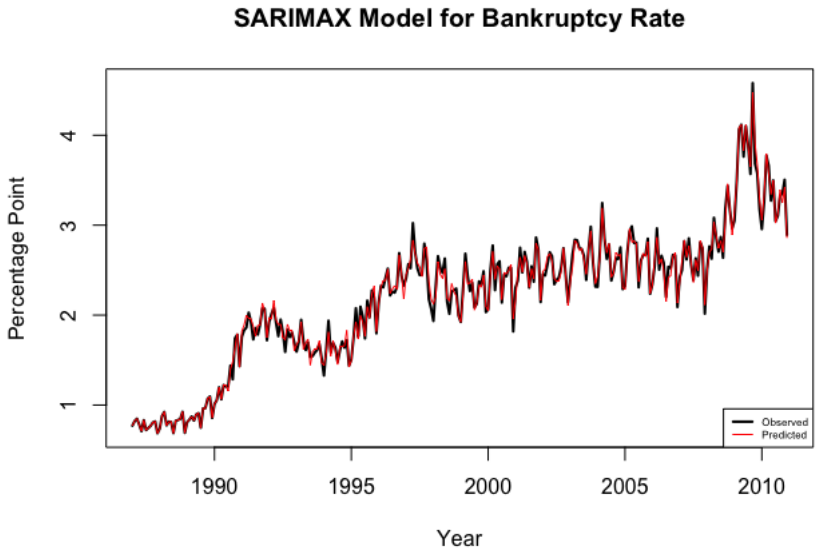

Figure 3: Best SARIMAX model performance on training set

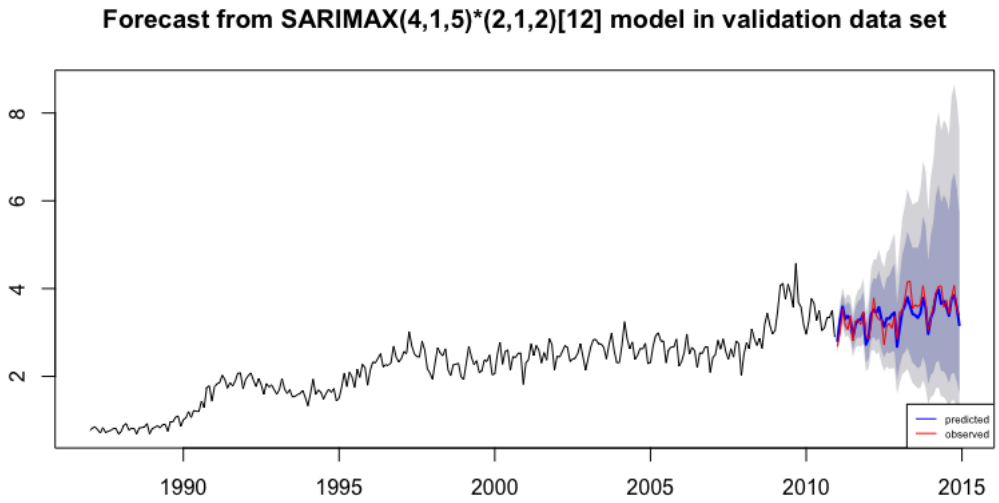

Figure 4: Best SARIMAX model performance on validation set

4.2 Model Assumptions

As previously already stated, SARIMAX(4, 1, 5)(2, 1, 2) is our best model based on the RMSE. Before fitting this model, we found that the response variable, bankruptcy rate, is not normally distributed, even after the Box-Cox transformation. Therefore, we applied the conditional-sum-of-squares method to fit the model, which does not require the response variables to be normally distributed, instead of the maximum likelihood method. After fitting a SARIMAX model, there are three model assumptions we need to check:

Metric Table		
Model	RMSE	Variance
Box-Jenkins	.229	0.0046
Holt-Winters	.276	0.1767
SARIMAX(4, 1, 5)*(2, 1, 2)	.206	0.0048
VAR(2)	.297	Does not apply

Table 1: Metric Table

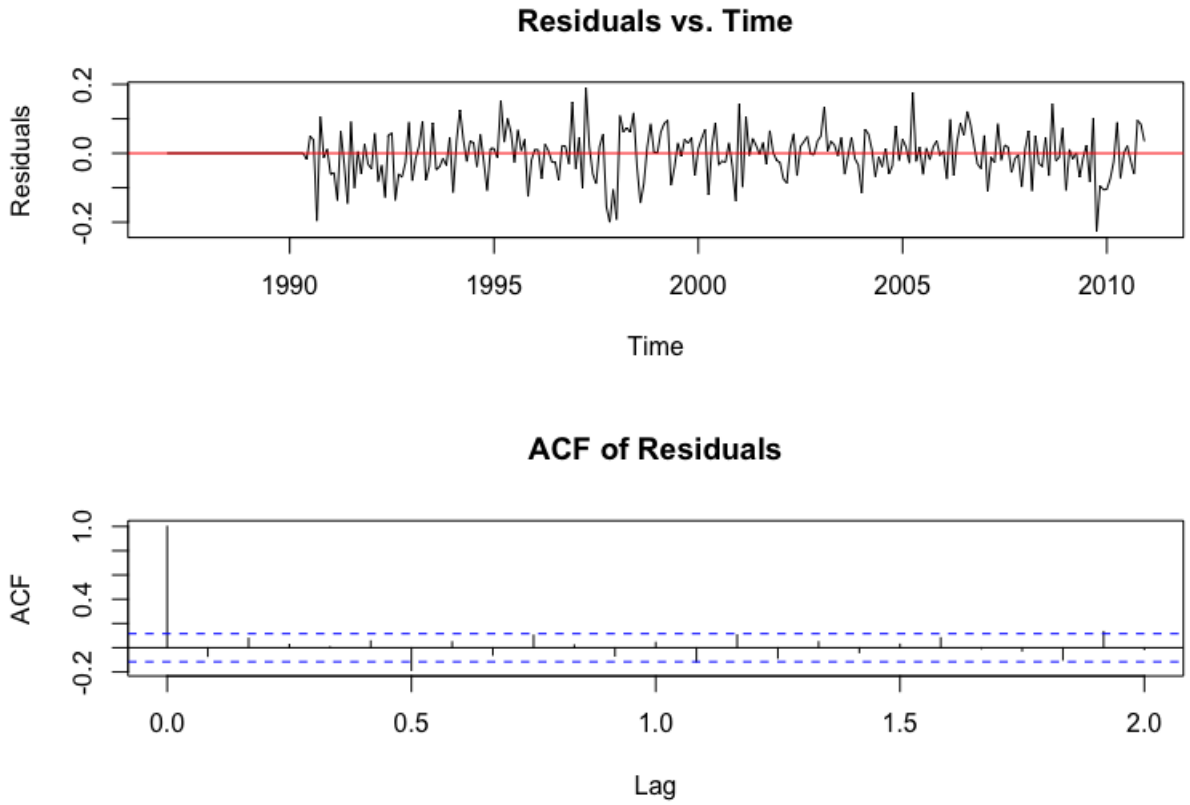

Figure 5: Model Assumption Test Plots

4.2.1 Zero-Mean

The zero mean assumption requires the mean of the residuals of this model to be zero. From the residuals vs time plot in Figure 5, we can see that the residuals are roughly symmetric around red line at 0, indicating the mean of the residuals is around 0.

4.2.2 Homogeneity

The homogeneity assumption states that the residuals of the model should be homogeneous. As we can see from the residuals vs time plot, the magnitudes of the residuals' fluctuation remain constant. There is no funnel shape, which would indicate heterogeneity in the residuals.

4.2.3 Uncorrelated residuals

Uncorrelated residuals requires the residuals of the model to have no self-correlation. In order to verify this assumption in our model, we plotted the auto-correlation function (ACF) from lag 0 to lag 24. There are no

spikes after lag 0, suggesting that the residuals at lag 0 are not correlated to its previous values. The uncorrelated residuals assumption are met.

In summary, our SARIMAX(4, 1, 5)(2, 1, 2) satisfies all three required assumptions: zero mean, homogeneity and uncorrelated residuals. Therefore, this model is selected as our best model that achieves the lowest RMSE and meets all assumptions.

5 Forecast Results

In order to forecast the bankruptcy rate from 2015 to 2017, we re-trained our best model on the entire training data set (from 1987 to 2014). The forecasting bankruptcy rate based on this model is plotted in Figure 6. The blue line represents the mean of the prediction. The shadows represent different prediction intervals, which are estimates of an interval in which a future observation will fall with a certain probability. The dark shadow represents the 80% prediction interval and the light shadow represents the 95% prediction interval. Based on our prediction results, the mean value of personal bankruptcy rate ranges from 3 to a little above 4. The bankruptcy rate is relative stable compared with previous years, which contributes to economic stability in Canada. According to our prediction intervals, with 95% probability, the personal bankruptcy rates from 2015 to 2017 ranges from 3 to 5. The bankruptcy rate will not exceed 5 with 95% probability, which is a good signal for the economy in Canada.

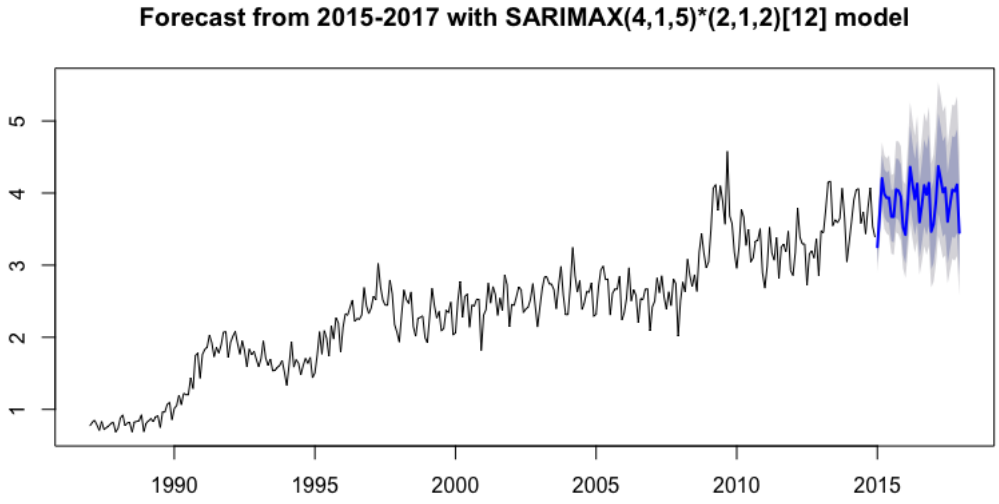

Figure 6: Forecasting from 2015 to 2017

6 Conclusion

To conclude: our best performing model was the SARIMAX(4,1,5)(2,1,2) model. This model gave us the lowest RMSE on our validation set. Our SARIMAX model predicts that bankruptcy rate in Canada during 2015-2017 will increase overall but not by very much, while still exhibiting seasonality. Though SARIMAX does not take into account the likely endogeneity of the house-price index and unemployment rate variables, it outperformed the VAR model. We may then conclude that the influence of the response variable on the explanatory variables may not be enough to justify usage of a VAR model.

7 Legal Disclaimer

This report is written in United State of America and do not subject the federal law in Canada. There is no French version provided.