Forecasting Bankruptcy Rates

Valerie Amoroso & Lin Chen

December 10, 2016

Contents

Introduction

One of the main interests of businesses across the world is to be able to accurately predict into the future. A business owner may want to predict what their sales will be for the next month so they can plan appropriate expenditures. A Wall Street analyst wishes to predict the rise and fall of stock prices. A weatherman hopes to predict when the next storm will occur and where it will fall, to ensure the safety of communities. All of these applications involve looking at data and trends from the past, and using that past data to make predictions into the future. When a set of data depends on time, it forms what is called a time series, and the science of making predictions and estimations into the future is known as forecasting.

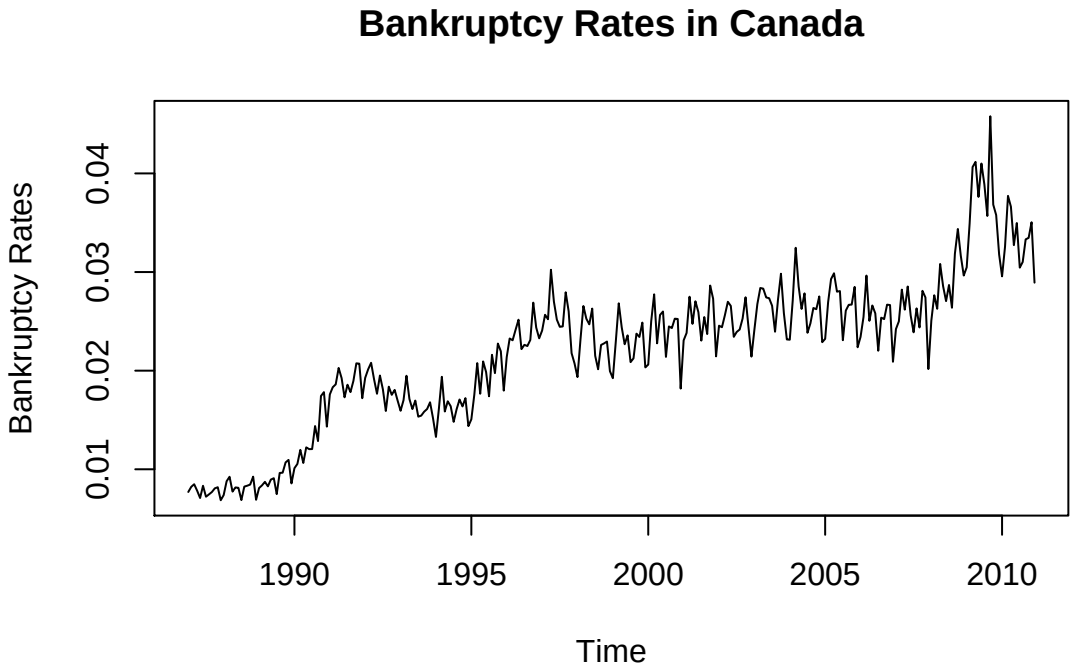
The problem we are dealing with at hand is to precisely and accurately forecast bankrupty rates for Canada. To do this, we are given four data points to consider: Unemployment Rate, Bankruptcy Rate, Population, and Housing Price Index. The unemployment rate is a measure of the prevalence of unemployment, calculated by dividing the number of unemployed individuals by all individuals currently in the labor force. The bankruptcy rate, which is the variable we wish to predict, refers to the rate of people who cannot repay the debts they owe to creditors and have had to file for bankruptcy. Population is the current number of people living in Canada, and housing price index measures the price changes of residential housing during a given time.

For each of these variables, we have monthly data from January 1987 to December 2010. Our goal is to use this data to create a time series model that will allow us to forcast into the future or make precise predictions about what the bankruptcy rates in Canada will be from January 2011 to December 2012.

Available Methods

In order to forecast into the future, we need to first create an appropriate model that describes the inherent structure of the time series. When creating a model, there are many available methods to consider. Typically as modelers, one of the first things to be considered is whether or not the raw data exhibits trend and seasonality. In this case, we say that a trend exists when there is a long-term increase or decrease in the data. Seasonality refers to whether or not a seasonal pattern exists within a time series, i.e. is it influenced by seasonal factors such as the quarter or the year, the month, or the day of the week. Seasonality is always of a fixed or known period.

Looking at an initial plot of bankruptcy data from from January 1987 - December 2010 it is appears that both trend and seasonality exist in our variable of question.



Since trend and seasonality both seem to exist in the series in question, the following modeling approaches are then most common to explore: Classical Decomposition, SARIMA, SARIMAX, Holt-Winters, VAR.

The idea behind each of these approaches will be described in more detail.

Classical Decomposition Approach

The goal of a decomposition approach is to construct from a given time series a number of component series where each of these components has a certain characteristic or type of behavior. In our case, we have already seen that trend and seasonality are behaviors that exist in our series, so we could use these as components, and fit a model including trend and seasonality and associated errors as components of our data. This is one of the most basic approaches in modeling a time series.

Holt-Winters

Another common approach to modeling a time series is an Exponential Smoothing approach, called the Holt-Winters method. Exponential smoothing works to smoothen data by applying what can be thought of as a filter that helps eliminate the "noise" in the way of error that is present in a time series. When trend and seasonality seem to exist, as they do in our case, we use triple exponential smoothing to model the time

series. This triple exponential smoothing involves a set of recursive equations and parameters alpha, beta, and gamma which can be chosen or tuned according to the level of smoothing desired at each level.

Before moving on to the next method, it is important to discuss a few more modeling terms for clarity. Let's first start with stationarity and differencing. Recall mean refers to the average value of a set of data points and covariance is a measures of how much two variables vary together. We say that a time series is stationary when its statistical properites such as mean and covariance are all constant over time, i.e., they are independent of time. Differencing refers to taking the value of a term at time t, and subtracting the value of that term one time point before (t-1). Differencing can be done for both trend, called ordinary differencing, and seasonality, called seasonal differencing. Both are integral parts to upcoming approaches.

In order to bulid a model with any kind of accuracy and precision, we require the assumption that something doesn't vary with time. For this reason, we have a class of models that we count on to adequately describe stationary time series', the most common being ARMA(p,q).

An ARMA model is actually the combination of an AR(p) and MA(q) models. An AR(p) model essesentially uses the past p observations to predict today's observations. An MA(q) model uses the past q errors to predict today's observation. Combining these we arrive at the ARMA(p,q) model that uses the past p observations and q associated errors to predict today's observation. So in our problem situation, if our original bankruptcy time series was stationary, we could use an ARMA(p,q) model to tell us how future bankruptcy rates are dependent on the p previous months bankruptcy rates and the past q errors associated with those rates. However, in our case and in many problem cases, the time series in question will not initially be stationary, and thus will need to undergo a transformation to become stationary. This is where differencing comes into play and brings us to the next common method of modeling.

SARIMA

We use a SARIMA model when trend and seasonlity are present in a time series. A SARIMA model is an ARMA(p,q) model that takes into account the need to 'ordinary' difference for trend d times, and seasonal difference for seasonality D times. Giving us a SARIMA(p,d,q)X(P,D,Q) model. Where P and Q in this case indicate how future bankruptcy rates are dependent on P previous seasons bankruptcy rates and the past Q seasonal errors associated with those rates. Once we difference d times for trend and D times for seasonality, we are left with a stationary time series that can then be modeled with an ARMA model as described above.

SARIMAX

Until now, the models that have been described have been univariate times series; that is, we have only considered using previous bankruptcy rates to make predictions of future bankruptcy rates. However, it is often possible that there are other variables present that are related with the one of primary interest (which we will call the response).

This may be the case in our problem, as we have been given three other variables to consider when trying to predict bankruptcy rates. Recall these include population, housing price index, and unemployment rate.

There are two ways that we can treat these variables, as exogenous, or as endodgenous. With exogenous variables, we say that these variables influence the primary variable of interest, the response, but the response is not influenced by these variables. With endodgenous variables, the external variables influence the response and the response also influences the external variables as well.

A SARIMAX model is a model that includes and considers these external variables as exogenous variables, and thus takes into account the relationship that exists between the primary response variable and the external variables.

Since we have data for three external variables, it would be good practice to try and fit a a SARIMAX model that takes into account that population, housing price index, and/or unemployment rate may have an influence on bankruptcy rate in Canada.

The SARIMAX model in particular still accounts for seasonality and trend and the need to difference to make the series stationary as SARIMA did.

The only limitation of SARIMAX is that it requires the future values of the exogenous variables. Fortunately in this project, we had all the 2011-2012 data for exogenous variables. If we do not have these values, the prediction interval would be too narrow.

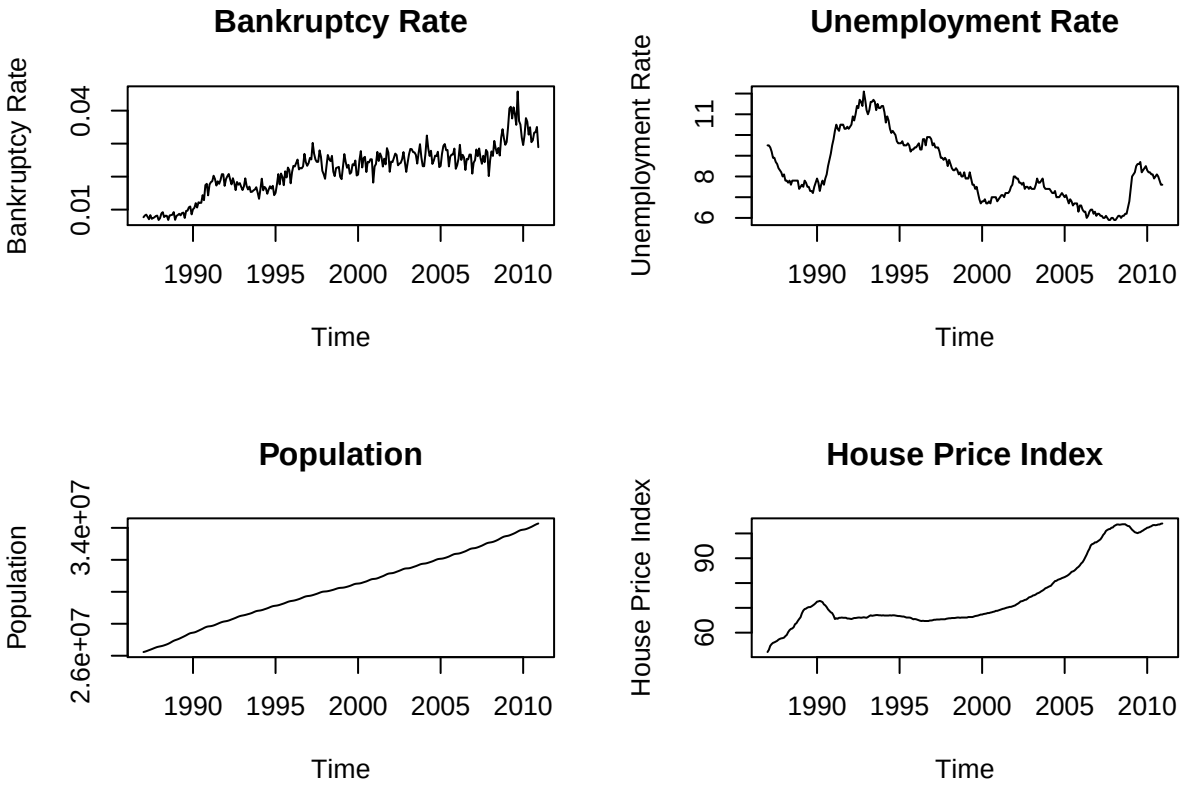
VAR (Vector Autoregression)

Finally, similar to SARIMAX is the VAR method of modeling. The difference between SARIMAX and VAR is that in VAR we treat all variables as endogenous - hence we assume that bankruptcy rates are affected by population, housing price index, and unemployment rate, AND that bankruptcy rate also has an influence on these variables as well. Thus the VAR method is good for capturing the inter-dependencies between the different time series models.

Our Method & Results

When searching for the best model to solve this problem, we tried using all of the methods listed above. In each case we first split the data into two sections, a training set and a validation set. This is a common practice when trying to predict into the future, as there is no metric that can be used to know how close your predictions are, since the future observations have not yet occured. To remedy this, we keep the first 80% of the observed data as the training set, which is used to build our model, and the last 20% of our observed data becomes the validation set, which is used to measure how well the model was able to predict by comparing the predictions to actual observations. One of the most frequent metrics used in quantifying a models predictive power is called the Root Mean Squared Error, or RMSE. RMSE is a measure of the differences between the predicted values and observed values. Thus if the model that was created has good predictive power, we would expect the predictions to be close to the actual observations during that time period, and thus our RMSE would be low. In searching for the best model, we used RMSE as our main metric of comparision, also taking into consideration the variance and model complexity. We found the optimized model for each approach. The RMSE table and the prediction graph of each approach was attached at the Appendix.

After trying each approach, the model that we found to make the most accurate and precise predictions was a SARIMAX model. Recall that in a SARIMAX model, we acknowledge that external variables are influencing our primary variable of interest, though initially we are not sure which ones. Since we had data for three external variables, population, housing price index, and unemployment rate, we first plotted these variables along with bankruptcy rate to see if it appeared there was a relationship over time.



What we noticed was that it appeared that the trends of all three variables were related to the bankruptcy rate. That is, as unemployment rate dropped, bankruptcy rate rose, and as housing price index rose, bankruptcy rate rose. The only variable in question was population. While population also increased over time, this increase was very linear, so it was hard to distinguish if this rising population was actually affecting bankruptcy rate, or just happened to have a simiar trend over time.

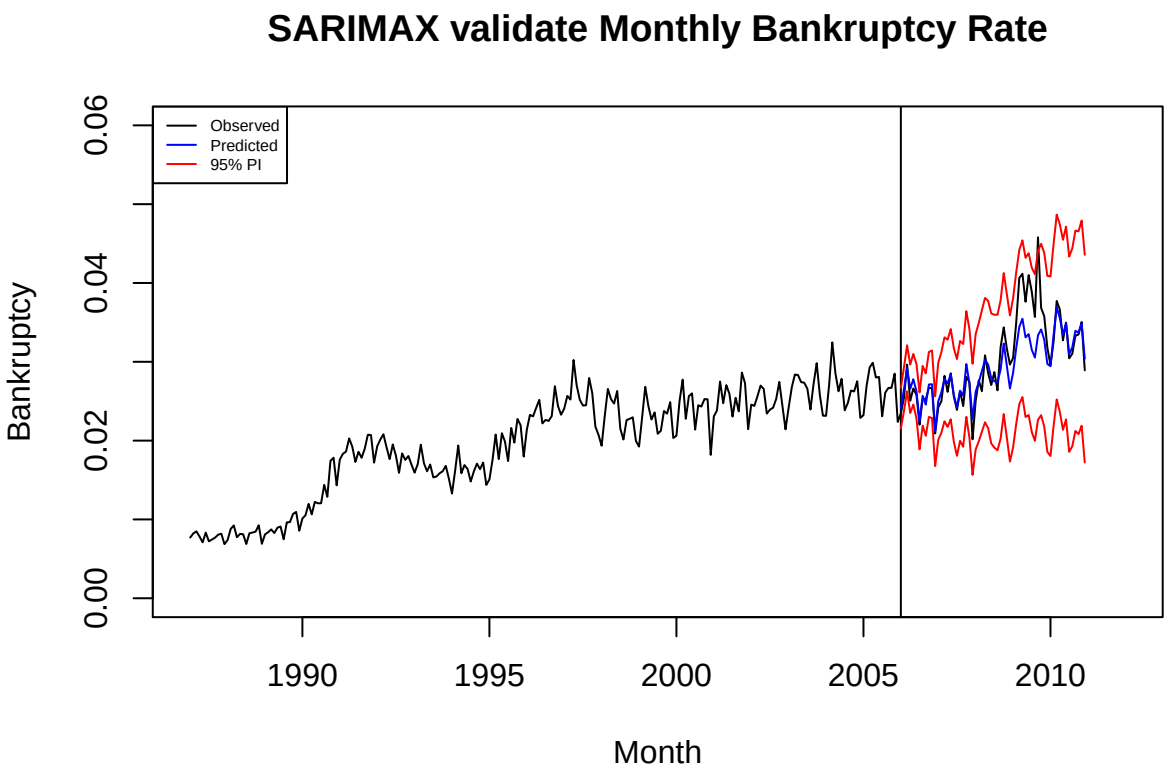
To evaluate whether or not this was the case, we created a variety of SARIMAX(p,d,q)x(P,D,Q) models, both including population as an external variable, and also not. We created a loop and cycled through a variety of models with different orders, where the order of a model refers to the values we will choose for p, P, q, Q, d, and D. We recorded the RMSE for each model and narrowed our choices down to the following two models.

The best model that we found which included the use of all three external variables, house price index, unemployment rate, and population, was SARIMAX(1,1,5)x(2,1,3). This model gave us an RMSE of 0.002840527.

The best model we found that did not include population, but still included house price index and unemployment rate was SARIMAX(4,1,5)x(5,1,3). This model gave us an RMSE of .002766579.

Our next step was to check the residual plots of these models. Residuals are the difference between the observed values and the predicted values, and by using the SARIMAX method we are making assuptions about the residuals that need to be checked. Both models met all assumptions and so we moved on to looking at the model complexity.

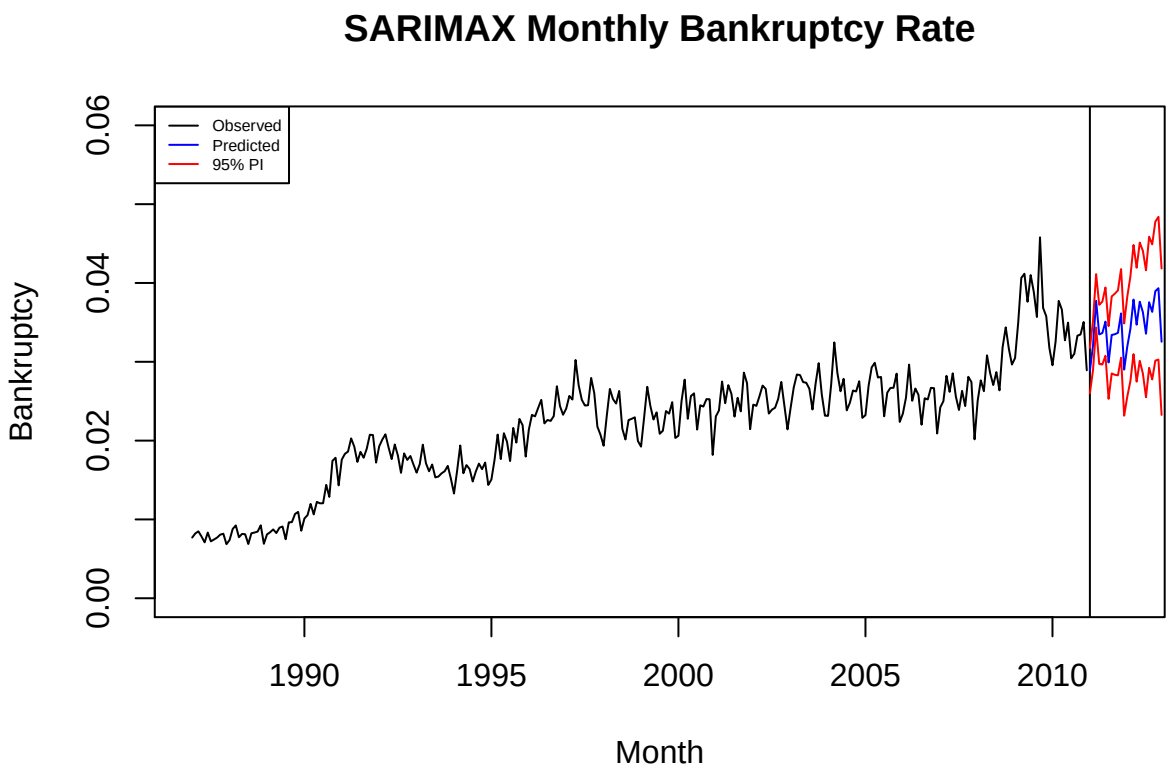
Comparing the orders of these two models it can be seen that the model that does not include population has higher values for both p and P. A value of $p = 4$ compared to $p = 1$, means that in order to make predictions into the future, we would need to use the previous four days worth of observations as opposed to one day. Thus, even though this second model not including population produced a slightly lower RMSE of .0027 as compared to .0028, we decided taking complexness into consideration the SARIMAX(1,1,5)x(2,1,3) is the best model. We feel this model not only produces accurate and precise predictions, but is also significantly simpler and easier to interpret. The results of our model's predictions against the validation set can be seen below.



We see that the actual observations of bankruptcy rates in Canada between 2006 and 2010 fall close to our predictions, and almost uniformly withing our 95% confidence bands. This result confirms our choice in our model.

Conclusion

With our model and order chosen, we are now able to use all of the original data on bankruptcy rates, from January 1987- December 2010, to forecast the January 2011 – December 2012 bankruptcy rates. The results of these predictions in tablular and graphical form are shown below.



month	prediction	lowerPI	upperPI
January 2011	0.028867842459135	0.0259945168197128	0.0317411680985572
February 2011	0.0320673610609278	0.0289517919370944	0.0351829301847612
March 2011	0.0377341861186667	0.034343992835452	0.0411243794018814
April 2011	0.0334783859952442	0.0297339789071702	0.0372227930833181
May 2011	0.0336436083678852	0.0296503295164369	0.0376368872193335
June 2011	0.0351001425809951	0.0307687445196148	0.0394315406423754
July 2011	0.0299119169141489	0.0252983305243931	0.0345255033039047
August 2011	0.0334092339036055	0.0285207860554942	0.0382976817517168
September 2011	0.0334964168338012	0.0283504225987899	0.0386424110688126
October 2011	0.033672607730208	0.028280564793056	0.0390646506673601
November 2011	0.0361367149784911	0.0305096316392945	0.0417637983176878
December 2011	0.0290101145118619	0.0231573722051339	0.0348628568185899
January 2012	0.0319240003987853	0.0255848138551096	0.0382631869424611
February 2012	0.0341888033068315	0.0275553739386979	0.0408222326749652
March 2012	0.0378986655677414	0.0309727479184458	0.044824583217037
April 2012	0.0347108320587252	0.0274839762406291	0.0419376878768213
May 2012	0.037626454890908	0.0301263418057078	0.0451265679761083
June 2012	0.0362968562382286	0.0285112820927813	0.044082430383676
July 2012	0.0335590002755312	0.0255045906366179	0.0416134099144445
August 2012	0.037556809641652	0.0292403992701824	0.0458732200131216
September 2012	0.0363351547390336	0.0277650672635702	0.0449052422144971
October 2012	0.0389657068165274	0.0301492250564949	0.04778218857656
November 2012	0.0393364642356145	0.0302801218175716	0.0483928066536574
December 2012	0.0325456561420319	0.0232555350809086	0.0418357772031551

Our model takes into account not only past bankruptcy rates in future predictions, but also the impact that an increase in population, housing price index, and unemployment have on bankruptcy. With our thorough

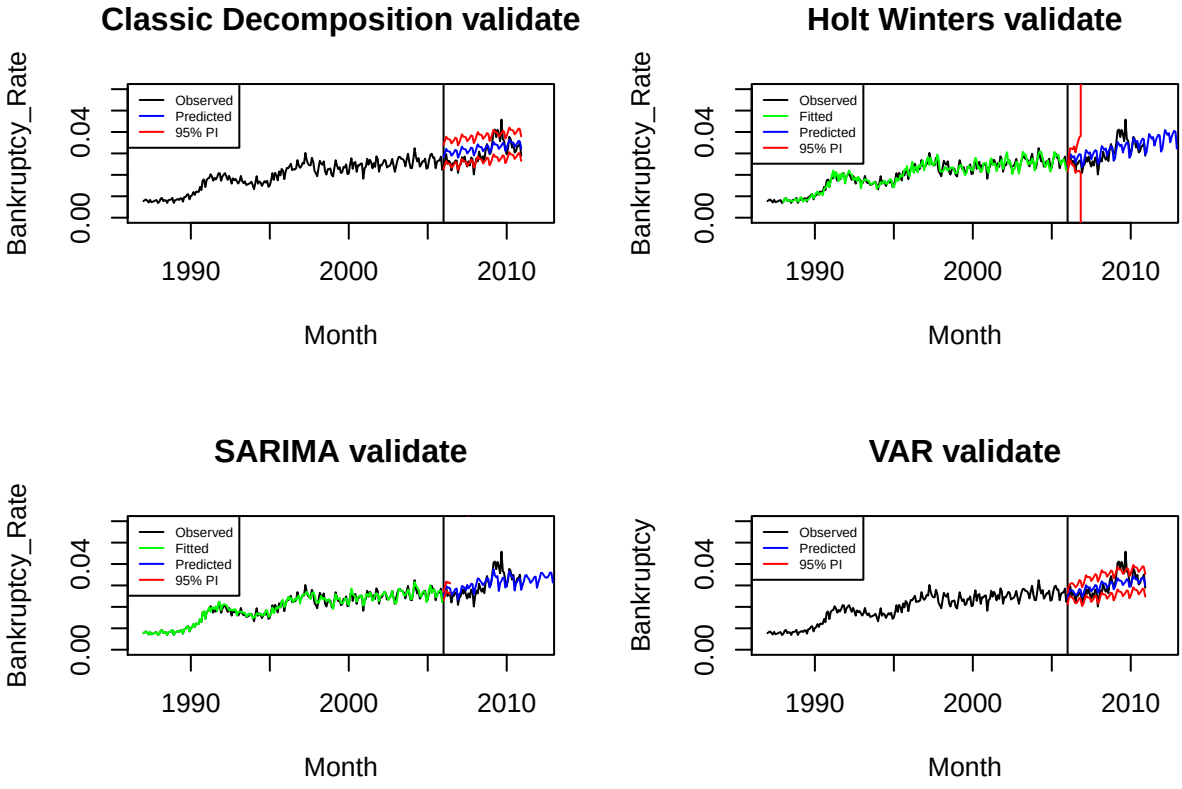
exploration through a variety of methods and models, we feel we have arrived at the model that creates precise and accurate predictions of future bankruptcy rates in Canada.

Technical Appendix

1. RMSE of validation set by using Classical Decomposition Approach, Holt Winter TES multiplicative $\alpha = 0.4, \beta = 0.3, \theta = 0.5$, SARIMA(1,1,1)x(3,1,5), VAR(p=5).

```
##                  Approach        RMSE
## 1 Classical Decomposition 0.004749869
## 2    Holt Winter TES mult 0.003712317
## 3   SARIMA(1,1,1)x(3,1,5) 0.003727968
## 4                VAR(p=5) 0.003822241
```

2. Prediction for validation set, using Classical Decomposition Approach, Holt Winter TES multiplicative $\alpha = 0.4, \beta = 0.3, \theta = 0.5$, SARIMA(1,1,1)x(3,1,5), VAR(p=5).



3. Prediction for test set, using Classical Decomposition Approach, Holt Winter TES multiplicative $\alpha = 0.4, \beta = 0.3, \theta = 0.5$, SARIMA(1,1,1)x(3,1,5), VAR(p=5).