Canadian Bankruptcy Rate Forecasting

Feiran Ji, Yue Lan, Akshay Tiwari, Yiqiang Zhao

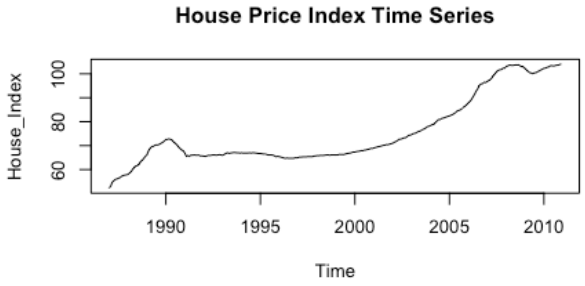
1. Introduction

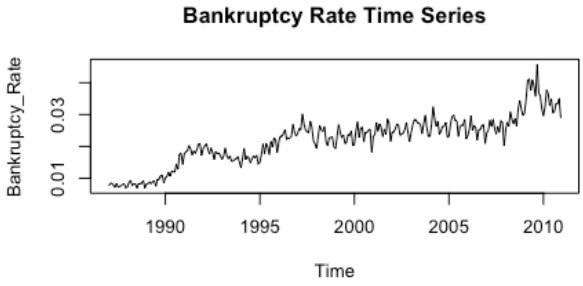
1.1 Problem Description

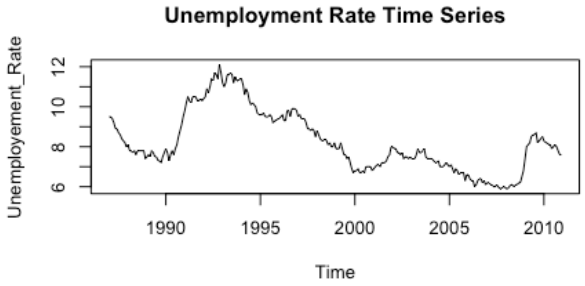
Bankruptcy in Canada allows people to erase debts when they are insolvent. It is a debt relief option of last resort. Understanding how the national bankruptcy rates change in time is an important task for risk management and is of interest to national banks, insurance companies, credit-lenders, politicians. In this report, based on monthly historical statistics of *bankruptcy rate*, as well as other related factors including *unemployment rate*, *population*, and *house price index* from 1987 to 2010, we want to forecast the *bankruptcy rate* in 2011 and 2012.

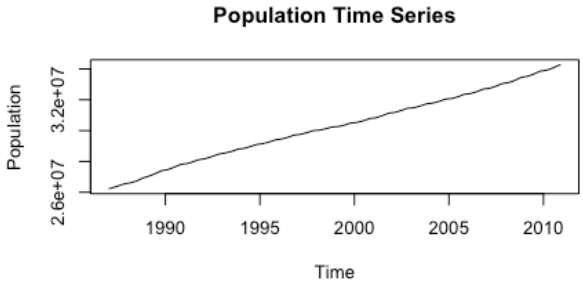
1.2 Exploratory Data Analysis

We first took a brief look at the 24-year time series of 4 variables in the dataset. From the plots we observed that *population* linearly increased overtime. *Unemployment rate* followed a decreasing trend while there were 3 bumps at 1991, 2002 and 2010. *House Price Index* went through a general increasing trend except for the temporary decrease at the year of 1991 and 2008. *Bankruptcy rate* followed a similar trend as with *house price index*. As all three variables exhibited abnormal patterns around year 1991 and 2008 when the financial crisis took place, we would need to pay special attention to these time periods since the market and statistics we are interested in are significantly affected by the crisis.

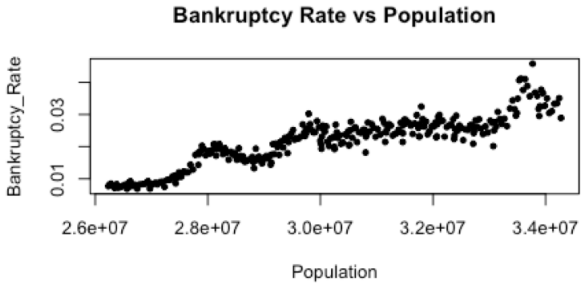


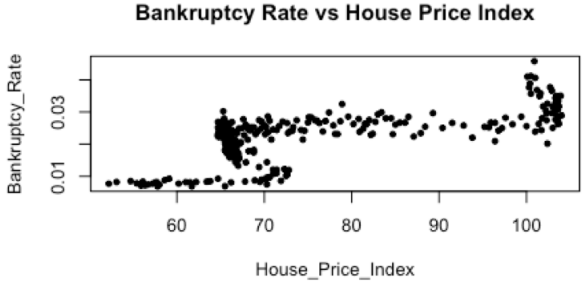






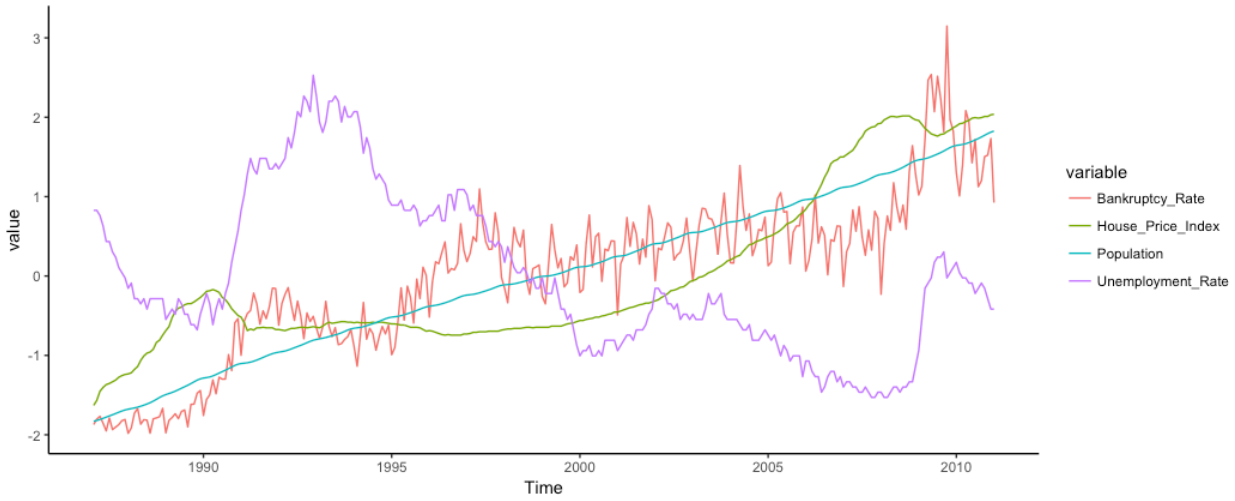
We also looked at the following pairwise correlation plots and coefficients for all the variables. Some of them had really high correlation with the target variable, *bankruptcy rate*, indicating that these variables may have influences on each other and thus playing a role in forecasting future values for *bankruptcy rate*.





	Unemployment Rate	Population	Bankruptcy Rate	House Price Index
Mean	8.236458	30,256,218	0.02190388	75.21806
Standard Deviation	1.527959	2,199,282	0.007590775	14.12351
Correlation				
Unemployment Rate	1.0000000	-0.5431182	-0.3169070	-0.5430593
Population	-0.5431182	1.0000000	0.8984050	0.8601513
Bankruptcy Rate	-0.3169070	0.8984050	1.0000000	0.6897080
House Price Index	-0.5430593	0.8601513	0.6897080	1.0000000

Moreover, we can compare the general trending of different variables by scaling the data and plotting them in a single graph. From the plot we can see that the trending of *bankruptcy rate* followed the trends of *unemployment rate* and *house price index* in a lagged pattern. We may consider including the lag of these variables when building the predictive model.



2. Modeling Approaches

As previously discussed, we are interested in the *bankruptcy rate* in Canada, which is part of a set of interdependent data including *population*, *unemployment rate* and *house price index*. Given that the target variable we want to predict is a time series and are influenced by other variables, we can choose from a variety of modeling approaches such as time series models, linear regression models, as well as other advanced models taking advantages of machine learning algorithms.

2.1 Time Series Models

2.1.1 Univariate Time Series

- ARIMA / SARIMA (Box-Jenkins Approach)

  SARIMA models a univariate time series by differencing the data to account for trend and seasonality and by fitting the data to an appropriate autoregressive and moving average model to account for other characteristics within the time series. If there is no seasonality present, it can be reduced to an ARIMA model.

By construction, the model directly takes into account trend, seasonality and most recent lags of the time series. Thus, the more recent the data point is, the greater impact it has on making predictions. SARIMA is therefore very helpful in time series predictions.

- Exponential Smoothing (Holt-Winters Approach)

  Exponential smoothing is a technique used to smooth time series data by iteratively assigning exponentially decreasing weights over time. With double exponential smoothing and triple exponential smoothing, we are able to further smooth the trend and seasonality in the data. We consider additive triple exponential smoothing (later referred to as Holt-Winters approach) here as the data has both trend and seasonality and the variance doesn't increase in a multiplicative fashion. Similar to Box-Jenkins approach, Holt-Winters approach performs well in time series predictions.

2.2.2 Multivariate Time Series

- ARIMAX / SARIMAX

  SARIMAX model can be thought of as an effective SARIMA model with exogenous explanatory variables taken into account. In such setting, not only is the time series modeled based on historical values, but also on other explanatory variables that can influence the target variable.

  SARIMAX can take advantage of the modeling and predicting power of SARIMA, and enhance the performance by adopting additional information from the known explanatory variables.

- VAR / VARX

  VAR model also takes external variables into account. Different from SARIMAX, VAR assumes that all the variables, known as endogenous variables, are interdependent on each other and symmetric with no distinction of explanatory variables and response variable. That is to say, VAR models and predicts all variables in the multivariate time series simultaneously. A VAR model can also take into account explanatory variables that don't depend on the endogenous variables but have an influence on them. In this case, it turns into a VARX model.

  Though VAR models are useful for dealing with multivariate time series, it may not be the most efficient in the case as we are particularly interested in one target variable and the other variables are already known.

2.2 Linear Regression Model

In the linear regression setting, we can regress the target variable against other explanatory variables to model the relationships between them. Thus, we are able to predict the target variable by applying the relationships we modeled to the explanatory variables given.

Linear regression model is especially useful in understanding and interpreting the interactions between explanatory variables and the target variable. It is also helpful in predicting future values when the target can be properly explained by the other variables.

2.3 Others

- Random Forest

  Random forest model is an ensemble of a number of decision trees. The idea behind is to classify the data points based on their feature values and predict the response value for a new data point by finding the historical data points that are most similar to it.

  The predictions made by random forest can be very accurate for new data with feature values that are within the range of historical data. However, the model doesn't perform well with new data with feature values that it hasn't seen before as it can't properly find the similar historical data points. Thus, random forest may not be appropriate since we are predicting into the future with a time range new to the model.

To summarize, we will fit the data with linear regression, SARIMA, Holt-Winters, SARIMAX, and VAR models respectively and justify their predictive power.

3. Modeling and model selection

3.1 Evaluation Metrics

Since our purpose is to predict the bankruptcy rate in the next two years, we would like to maximize the prediction accuracy, which is equivalent to minimizing the prediction errors. A commonly used evaluation metric RMSE (root of mean squared errors) is an appropriate measurement of prediction errors in this case.

We will split the data into training sets (1987-2008) versus validation sets (2009-2010), train the models with training dataset, and evaluate the model's' predictive power by their RMSE on validation dataset.

3.2 Feature Selection

For the univariate models, we are only considering the target variable *bankruptcy rate* by itself and its historical values.

In terms of the multivariate models, we can easily derive the *bankruptcy population* and *unemployment population* by multiplying *population* by *bankruptcy rate* and *unemployment rate* respectively. With these intermediate variables, the goodness-of-fit as well as predictive accuracy of certain models get improved. Moreover, as mentioned beforehand, we also take into account the 9-months lag values of *unemployment rate* and 7-months lag values of *house price index* as they can enhance the model's performance.

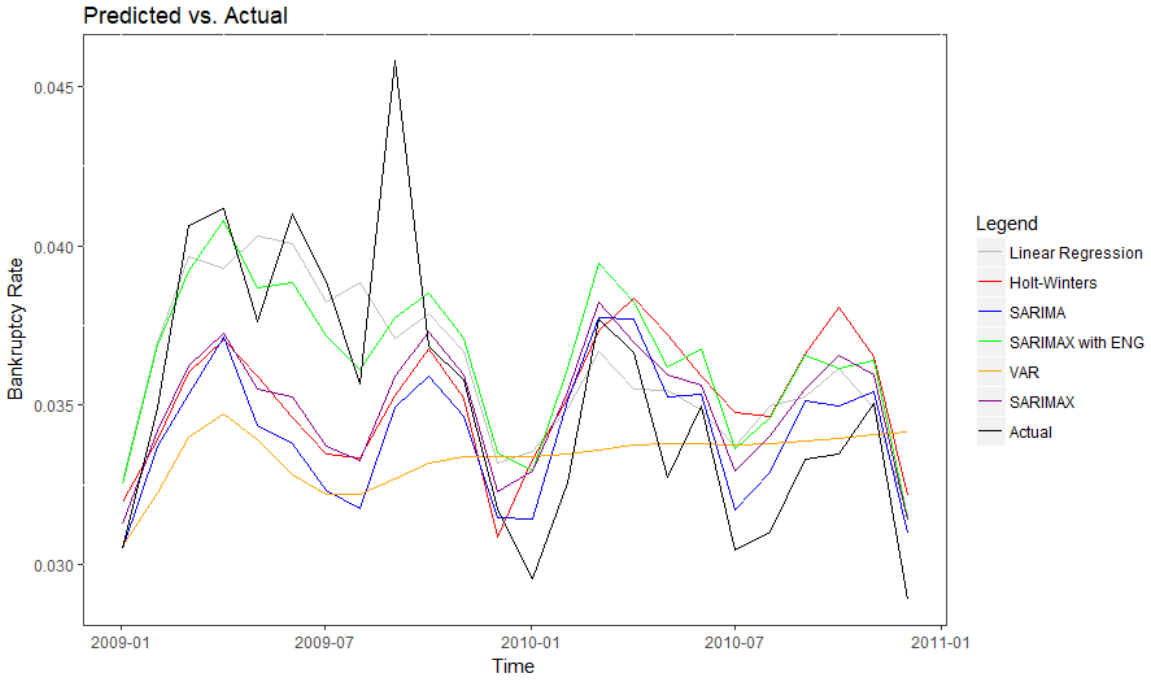
3.3 Model Selection

We iterated through different models in each approach to come up with the optimal models based on the selected evaluation metrics RMSE. Below are the optimal models under each approach and their respective RMSE, along with some other important performance metrics.

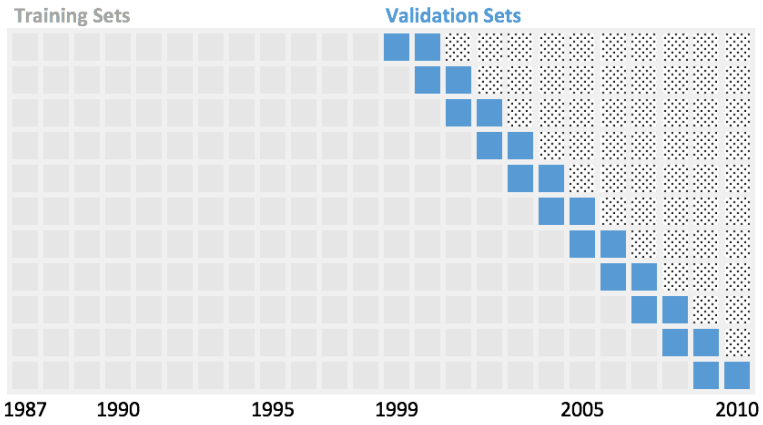
Optimal models under different modeling approaches
Training set: 1987-2008
Validation set: 2009-2010

Model	Log-Likelihood	AIC	RMSE (09-10)
Linear Regression	-	-	0.002787
SARIMA (1, 0, 3)(3, 1, 2)[12]	1,301.817	-2,583.634	0.003652
Holt Winters (α=0.66, β=0.01, γ=0.91)	-	-	0.003839
SARIMAX (3, 0, 0)(2, 1, 3)[12]	1,317.076	-2,610.153	0.003262
SARIMAX (3, 0, 0)(2, 1, 3)[12] (with feature engineering)	1,250.408	-2,472.815	0.002780
VAR (p = 4)	-	-	0.004628

A lower value of AIC and a higher value of Log-Likelihood indicate a better goodness-of-fit of the data, while a lower value of RMSE indicates a better predictive accuracy over the validation set. As our goal here is to predict the future values, we will focus on the predictive power.



However, the above RMSE values can only evaluate the prediction accuracy on the chosen validation set. To avoid overfitting, we adopt the cross-validation technique which takes several different subset of data as validation sets and evaluate the models by their average scores. The training sets and validation sets used are shown as follows.



The cross-validation RMSEs for each model are shown as follows.

Cross validation scores for different models
Training set: from 1987 to the beginning of each validation set
Validation set: all 2-year time windows from 1999 to 2010

Model	average RMSE over validation sets
Linear Regression	0.002378
SARIMA (1, 0, 3)(3, 1, 2)[12]	0.003624
Holt Winters (α=0.66, β=0.01, γ=0.91)	0.002941
SARIMAX (3, 0, 0)(2, 1, 3)[12]	0.002224
SARIMAX (3, 0, 0)(2, 1, 3)[12] (with feature engineering)	0.002080

Based on the above shown RMSEs, we selected ***SARIMAX (3,0,0) (3,1,1)[12] (with feature engineering)*** as the final model.

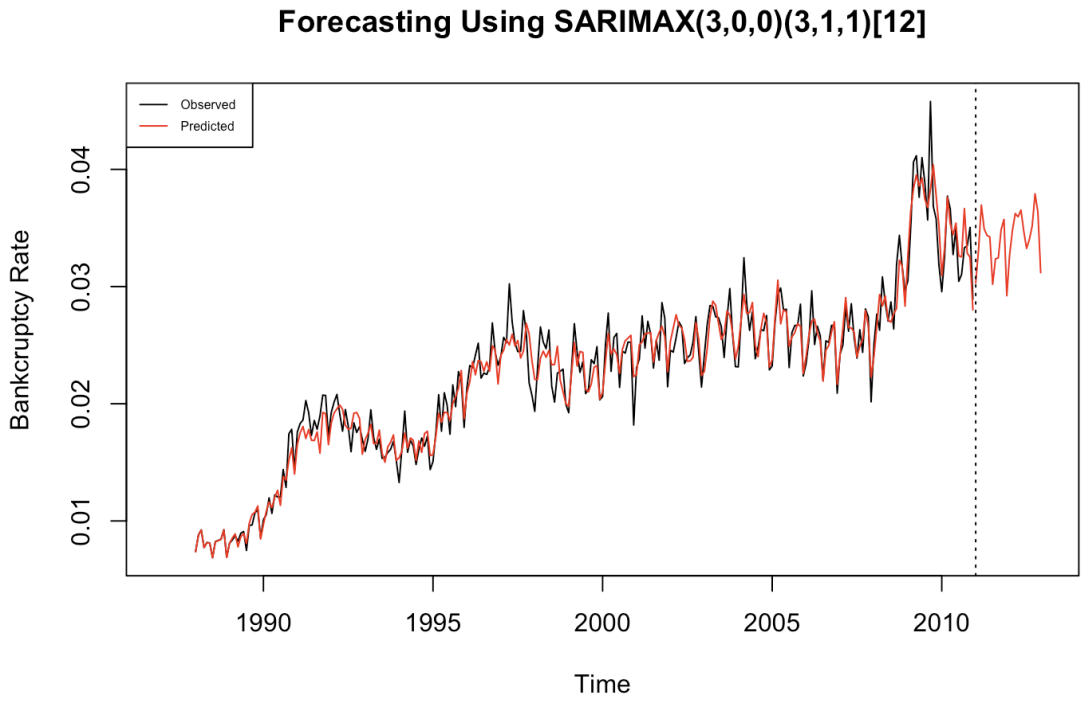
SARIMAX (3,0,0) (3,1,1)[12] (with feature engineering)

The final model takes into account the trend, seasonality and the influence of other variables mentioned above at the same time. To model the trend, it considers the historical data within the most recent 3 months. As for seasonality, it models the recursive pattern for each month across different years. Other variables and engineered features, as discussed above, have also been considered as exogenous variables to account for their influence.

4. Forcasting

With ***SARIMAX (3, 0, 0)(3, 1, 1)[12] with feature engineering*** model as our final model, we re-trained it using all data from 1987 to 2010 to make use of the latest information available. The following prediction values for the years 2011 and 2012 are given by the model.

Jan-11	Feb-11	Mar-11	Apr-11	May-11	Jun-11	Jul-11	Aug-11	Sep-11	Oct-11	Nov-11	Dec-11
0.03030	0.03311	0.03695	0.03492	0.03437	0.03423	0.03019	0.03237	0.03245	0.03486	0.03573	0.02922
Jan-12	**Feb-12**	**Mar-12**	**Apr-12**	**May-12**	**Jun-12**	**Jul-12**	**Aug-12**	**Sep-12**	**Oct-12**	**Nov-12**	**Dec-12**
0.03272	0.03479	0.03623	0.03595	0.03652	0.03473	0.03326	0.03404	0.03518	0.03792	0.03635	0.03119



5. Conclusion

We can conclude that as per our predictions the average monthly *bankruptcy rate* for the next two years is :

- 2011 : 0.03322486
- 2012 : 0.03490624

The general trend is upwards from 2010 to 2011 and 2011 to 2012, therefore the *bankruptcy rate* will increase from 2010 to 2011.