📉🇨🇦💸 Group 1.7: Forecasting Canadian Bankruptcy Rate

👨🏻👨🏻👨🏻👩🏻 Chong Geng, Evan Liu, Jialiang Shi, Katja Wittfoth

☝️ Description of the Problem:

The risk of bankruptcy is often priced into interest rates from the banks and other credit lenders or baked into the price for insurance rate from insurance companies. That is why knowing the national bankruptcy rate and being able to predict it is essential for many businesses.

In our project, we aim to accurately forecast the monthly Canadian national bankruptcy rates for the period from 2015 to 2017.

As a starting point, we have monthly data from January 1987 to December 2014 for Canadian consumer bankruptcy rate (%) as well as following monthly data for the period from January 1987 to December 2017:

- Unemployment Rate (%)
- Population
- Housing Price Index

We will show our modeling process and justify the selection of the final model and present our predictions.

👠 Modeling:

Times series is a sequence of observations each one being recorded or measured at specific times. In the case of bankruptcy rates, the observations were recorded monthly.

For modeling, we decided to split the data into a training set and validation set. We based our decision on following heuristics. First, based on the graph, we selected a proper time point to split our data so that we have an adequate amount of data points for training. By doing so, we also need to ensure that the pattern from the validation set does not completely differ from the training set. As a result, we found that a good benchmark is around **two to four years**. Second,

we wanted to have a similar amount of data points compared to the forecasted period. Therefore, we chose to have **two years** for validation, from January 2013 to December 2014.

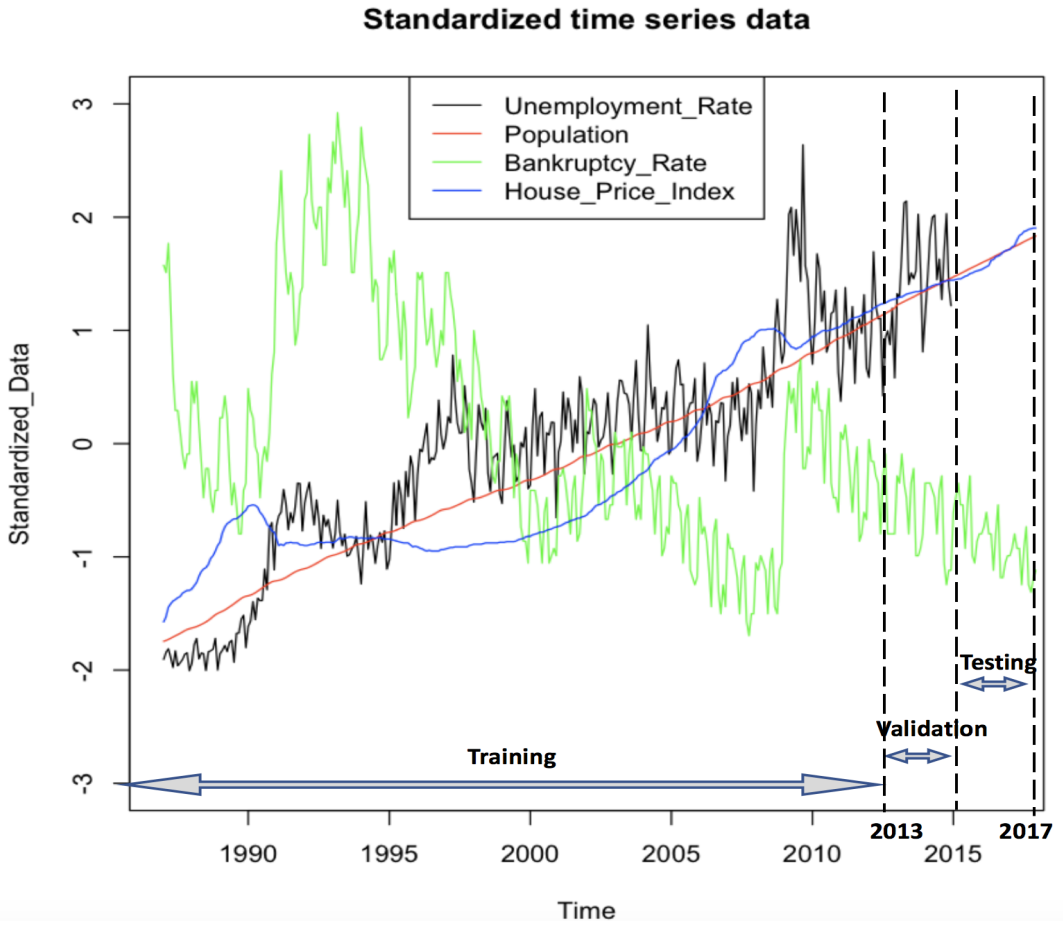

Chart 1: Standardized time series data split into Training/Validation/Testing parts.

Univariate Approach:

To do time series analysis, there are several methods we can use to model the data. If we want to model a time series based on its own history, this is called a univariate approach. There are several univariate approaches such as Exponential Smoothing models, in our analysis, we will only consider Seasonal Integrated AutoRegressive Moving Average model (SARIMA).

SARIMA methodology attempts to describe the movements in a time series as a function of what is called "autoregressive and moving average" parameters. These refer to AR parameters (autoregressive) and MA parameters (moving averages). This simply means that any given value can be explained by some function of its previous value, plus some unexplainable random error.

Multivariate Approach:

Unlike the univariate approach which models the time series based on its own history, the multivariate approach accounts for other variables that influence the prediction of the target time series.

For this project, we were provided with three additional variables, Unemployment Rate, Population and House Price Index. From an economics perspective, these variables may provide more help to accurately forecast the Bankruptcy Rate.

There are different multivariate models, such as SARIMAX, VAR, and VARX models. Compared to SARIMA, the SARIMAX model allow us to consider external variables, and it assumes that the external variables will influence the response time series, but the response time series will not influence the external variables (these external variables are exogenous). The VAR model, on the other hand, considers the external variables and the target time series to have an influence on each other (these external variables are endogenous). The VARX model, different with both two models, will assume some variables have a uni-direction effect on the time series, some would influence the modeled time series symmetrically.

In this case, it's hard to tell the relationship of Bankruptcy Rate with the other three variables are one- or two-direction, so we tried to build all three kinds of models and apply a metric to select the 'best' models in next section.

🛍️ Choice of model:

When it comes to how to choose the best model, we can measure the performance of our model from two perspectives: goodness of fit and predictive accuracy. By the goodness of fit, we mean how well our model fits the historical data. By predictive accuracy, we mean how well our model can predict the data in the future. In the context of our project, we are more concerned obtaining a more accurate future forecasting, so we will use predictive accuracy as the metric while selecting our model.

To be specific, we choose RMSE as our metric to evaluate the predicting performance of our model. RMSE stands for root mean squared error and represents the error rate between the prediction value and actual value. The smaller RMSE is, the more accurately our model forecasts the future values.

Before comparing the RMSE scores on the validation data of different models, we came up with the idea that applying “ensemble models” to enhance the predictive accuracy of our models. The idea was inspired by the “bagging” method of decision trees in the machine learning algorithm “random forest”: taking the top ensemble of “best” models that well match the training data then take an average of their predictions would be better than only take a single best model. Therefore, for each model (SARIMA, SARIMAX, VAR, and VARX), we applied a grid search in a reasonable model parameter space, then selected the top 10 models which had the best performance on the training model, and finally evaluated their performance on the validation data, the prediction RMSE were averaged and plotted below:

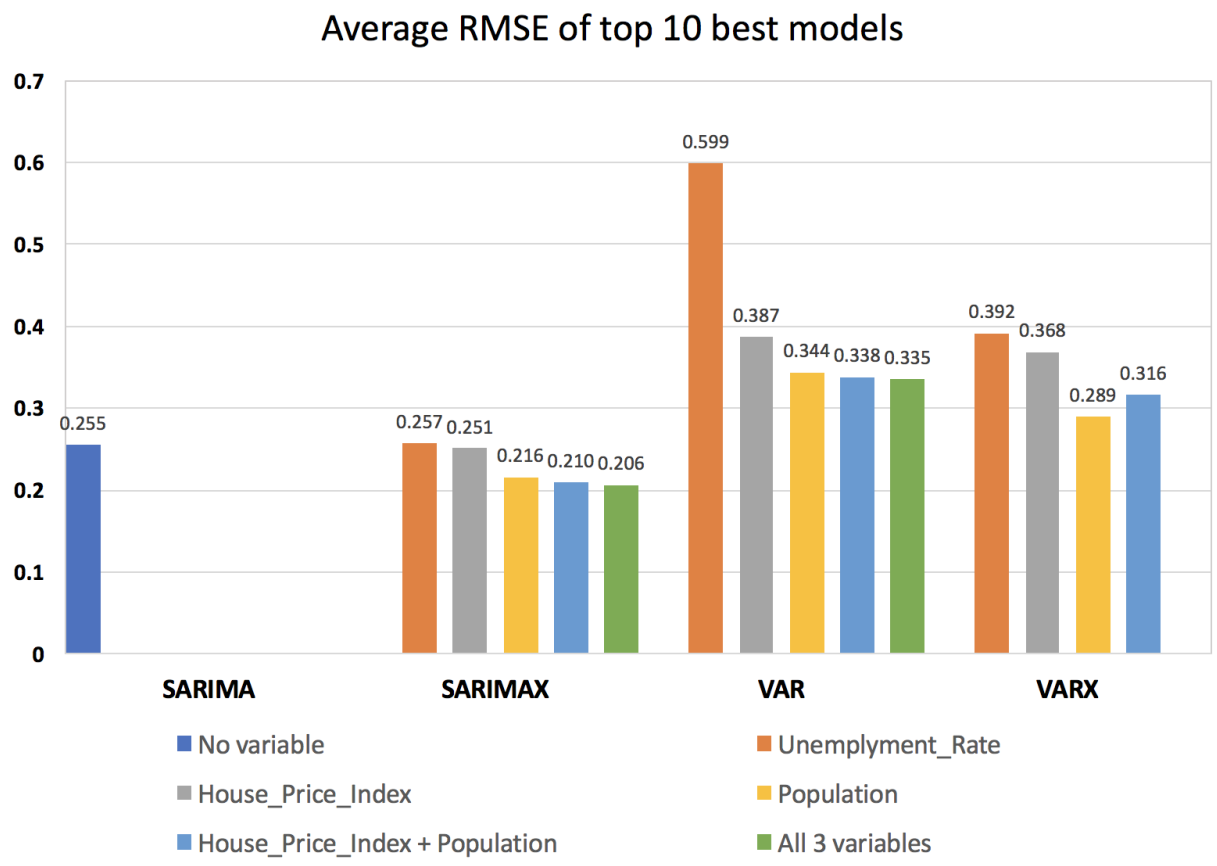

Chart 2: Average RMSE of top 10 best models

As mentioned before, the SARIMA model is univariate, so it included no additional variables and made a prediction based only on its own history; however, other three kinds of models are multivariate, so different plans were made to determine which variable should be considered when we made the prediction of bankruptcy rate. From the plot, we can tell that the SARIMAX model with all three extra variables (Unemployment Rate, Population and Housing Price Index)

gave the lowest RMSE score. Thus, we chose the top 10 best “SARIMAX” models to make our prediction.

We found other interesting insights from the above plot. Based on our modeling and analysis of predictive accuracy, VARX model does perform better than VAR model since VARX can handle both exogenous variables and endogenous variables. Moreover, while VAR model needs to predict every single endogenous variable based on historical data, VARX model is able to take advantage of the latest explanatory data. In our case, we regard the bankruptcy rate and housing price as endogenous variables, population, and the unemployment rate as exogenous variables. While VAR model predicts based on historical data, VARX model is able to take advantage of the latest population and unemployment rate to predict the bankruptcy rate of the same period.

In our case, VARX and SARIMAX models performed better than VAR and SARIMA models since they were able to take advantage of the explanatory variables in the test dataset. However, if we were to apply SARIMAX and VARX models to predict in real life, we simply are not able to collect those corresponding explanatory variables in the future. In other words, we need to predict those explanatory variables at the same time, which is not an easy task.

🌤 Forecasting:

Below you can see our prediction of Canadian bankruptcy rate for the period January 2015 till December 2017 using our ensemble SARIMAX models which are using Unemployment Rate, Population and Housing Price Index as exogenous variables:

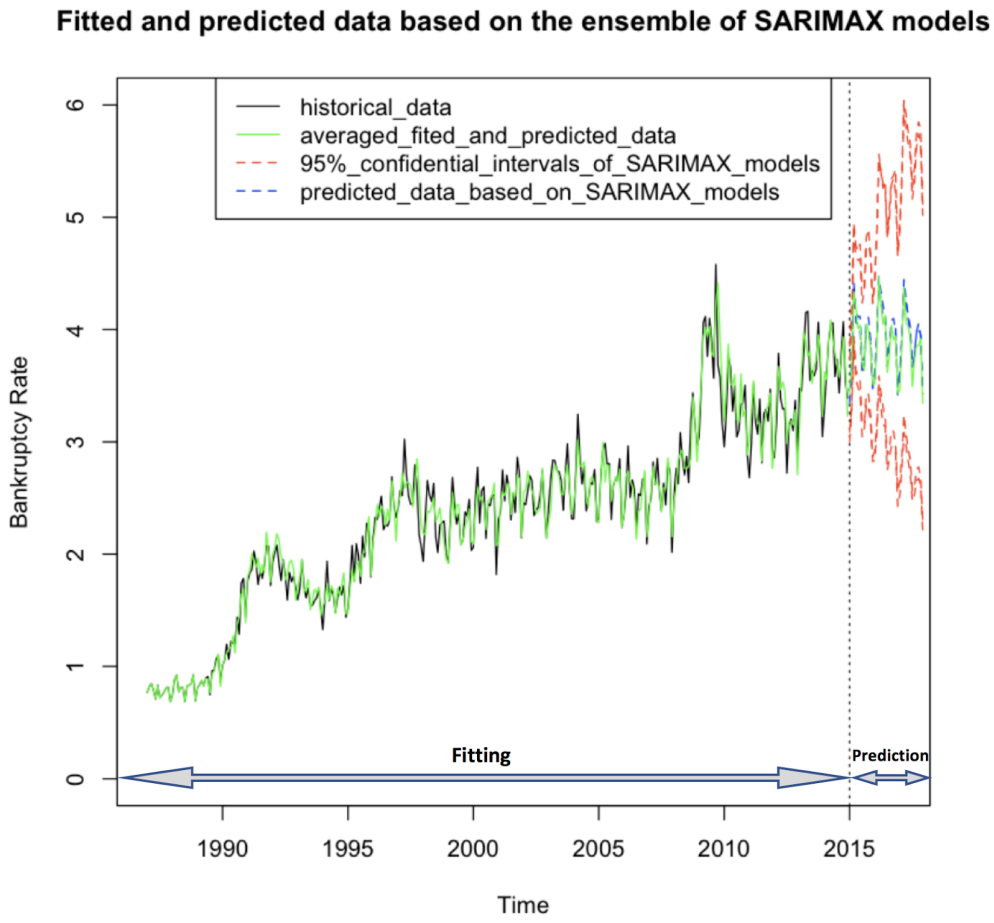

Chart 3: Fitted and predicted Canadian bankruptcy rate.

The tabular forecasting results are presented in the following table:

prediction	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2015	3.4308	3.7023	4.3384	4.1953	4.0090	4.0391	3.6964	3.6337	4.0489	3.9895	3.8403	3.5121
2016	3.5151	3.8409	4.4730	4.2292	4.0462	4.1263	3.6177	3.7690	4.0370	3.9099	3.9418	3.4321
2017	3.4642	3.8507	4.3721	4.0704	4.0219	3.9342	3.4936	3.7206	3.8563	3.8600	3.9216	3.3423

Table 1: Forecasted monthly Canadian bankruptcy rate for 2015 - 2017 with ensembled SARIMAX model.

It has to be mentioned that the fitted value and predicted value are based on the average of the ensemble of top 10 models; the red dash lines show the 95% prediction intervals of each model. The predictions in a single column are enclosed in a separate .txt file.