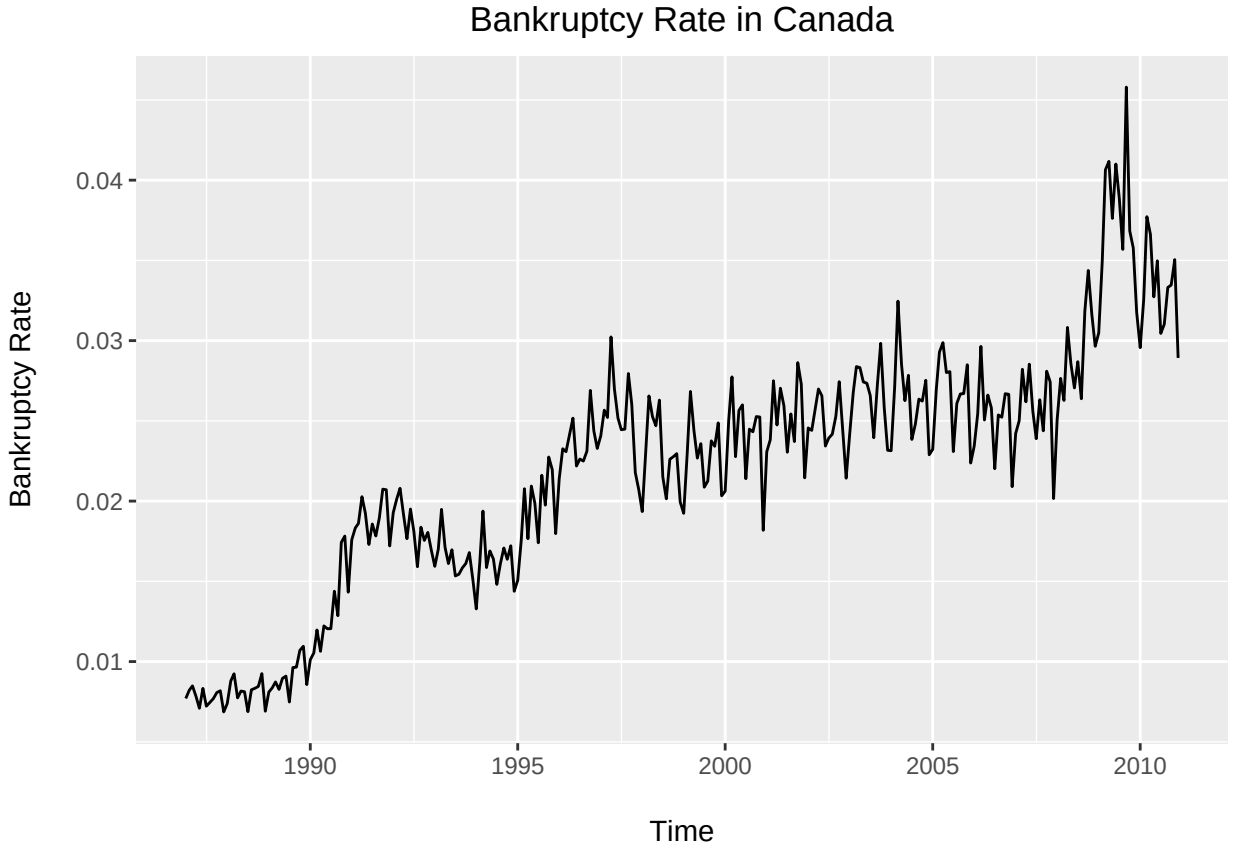
MSAN 604 Final Project

Spencer Stanley, Nishan Madawanarachchi, Vinay Patlolla, Jingjue Wang, Fang Wang

12/7/2017

Problem Statement

The goal of this report is, given information from the prior 21 years, to forecast the bankruptcy rate in Canada in 2011 and 2012. The dataset with which we worked contains month-by-month observations of the unemployment rate (as a percentage), population (in pure numbers), bankruptcy rate (as a raw rate, not a percentage), and the housing price index (in its natural scale). For reference, the bankruptcy rate over time is plotted below:



Modeling Approach

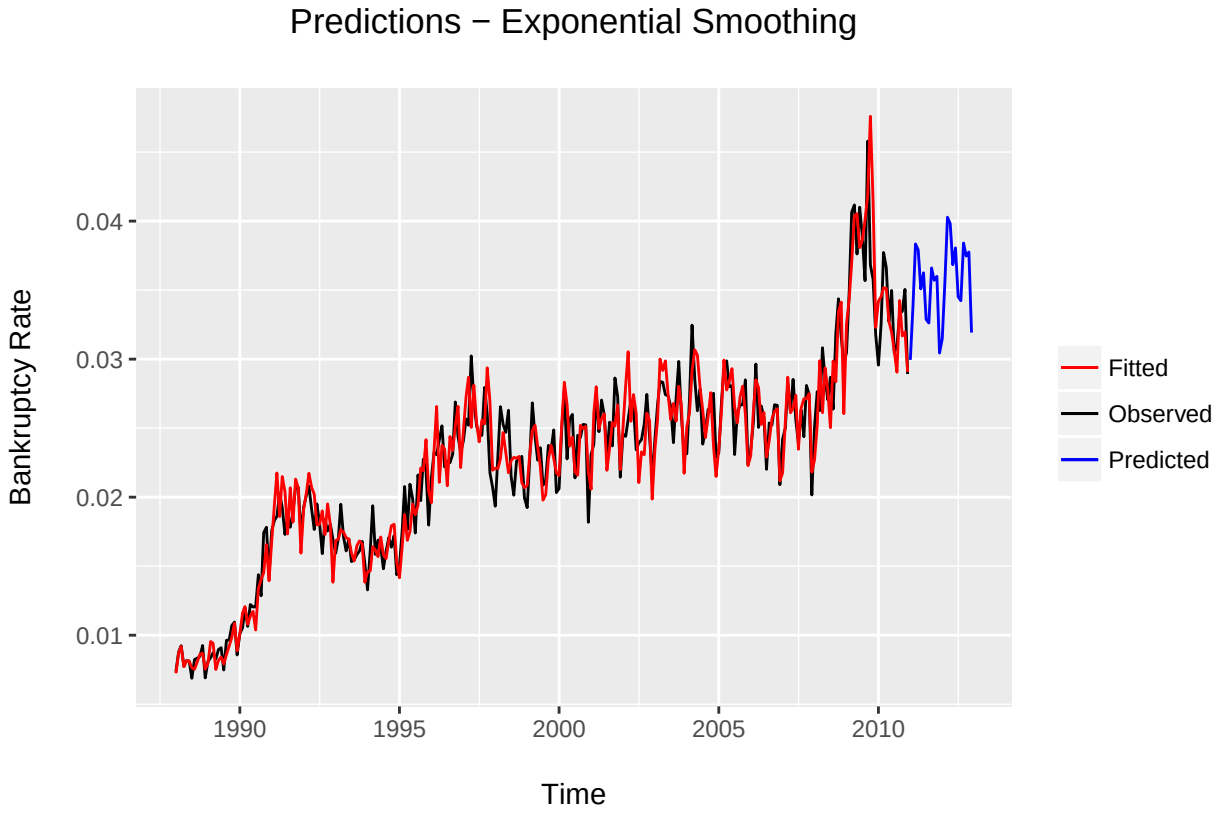
The choice of a validation set for this data is nontrivial. Conventionally, we would select a validation set either by pulling an interval the same duration as the test set (2 years) or by splitting the data approximately along some threshold like 80% train, 20% validation. However, it is important to note that the bankruptcy rate in 2008-2009 is inflated dramatically by the financial crisis. If that period were to be in our training set, we would likely appear to miss the mark on the ensuing recovery. On the other hand, if the crisis were part of the validation set, then we would be attempting to fit explicitly to this unforseen exogenous variable. In the end, our solution to this was to train our models on the period from 1987 to 2005 and validate them on the 2006-2007 period. This gives us a validation set which mirrors the test set both in terms of timespan (2 years) and in terms of relative financial stability.

We evaluated our models by way of the root mean square error (RMSE) on the above validation set. The models we worked with are detailed below.

Holt-Winters (Exponential Smoothing)

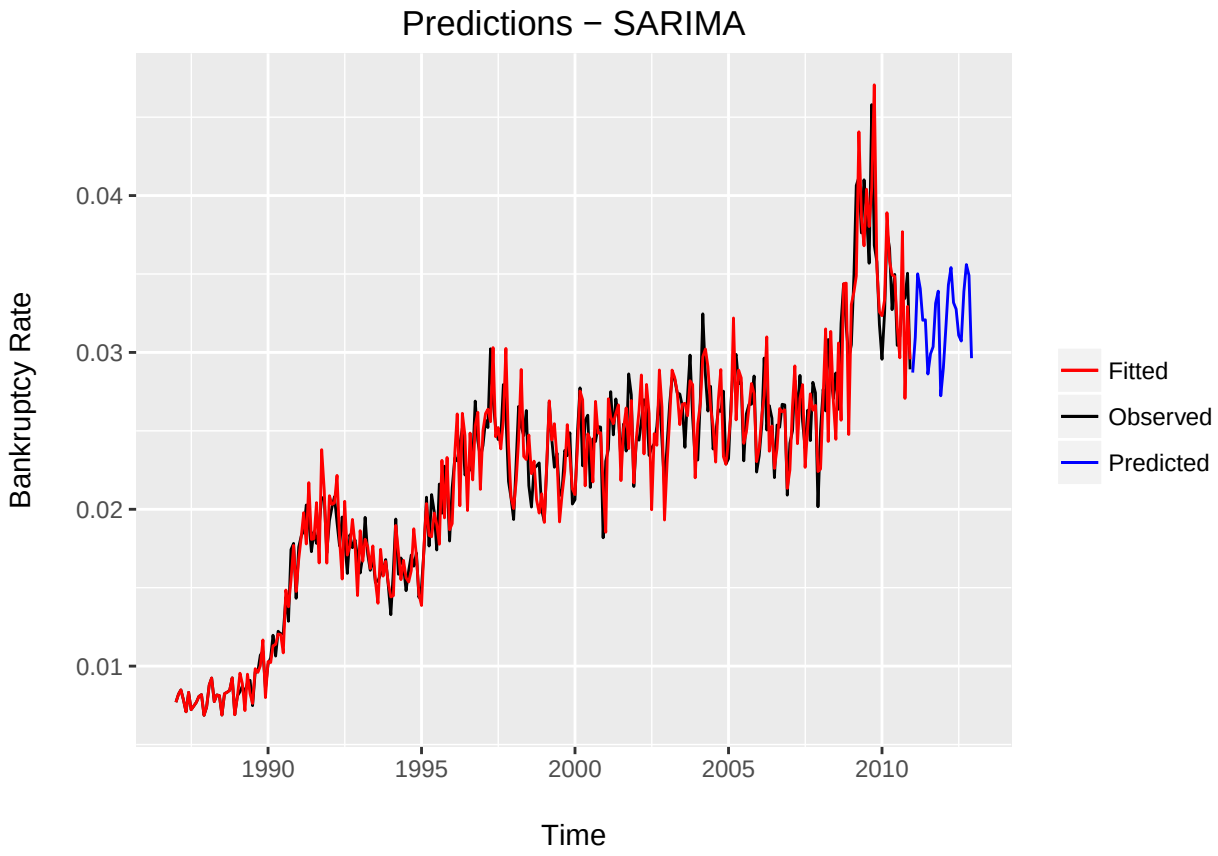
Exponential smoothing is a method of time series modeling by which we model the level, trend, and seasonality (if present) of a given time series by an exponential equation for each. Since our data exhibits both trend and seasonal components, we employ triple exponential smoothing to forecast the data. We can model this seasonal component either additively or multiplicatively, depending on how the variation in our data changes over time. Since that variation does appear to inflate over time (in a non-linear sense), we selected multiplicative seasonality for our model.

The parameters α, β, and γ, representing the model's sensitivity to the level, trend, and seasonality, respectively, were found using a grid search through potential values and selected the combination with the lowest MSPE.

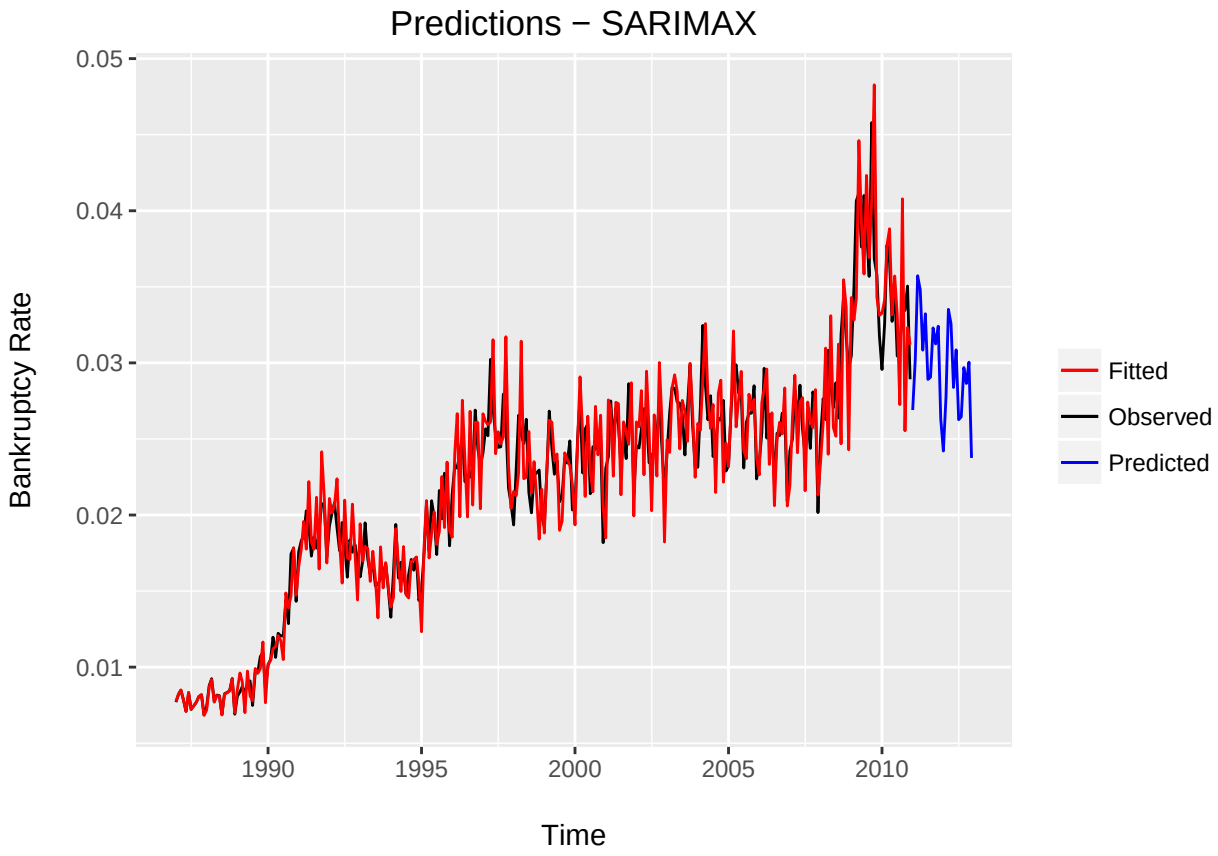


SARIMA / SARIMAX

SARIMA modeling involves stripping the data of its trend and seasonal components and then modeling the resulting "stationary" time series. Once we have removed the trend and seasonality, we select parameters to control the autoregressive and moving-average components of our now-stationary time series (and similar parameters for the seasonal aspects of the same). It is worth noting that SARIMA is a univariate model; that is, this model depends only on prior observed values of the bankruptcy rate and ignores the observations of other variables from the original dataset.



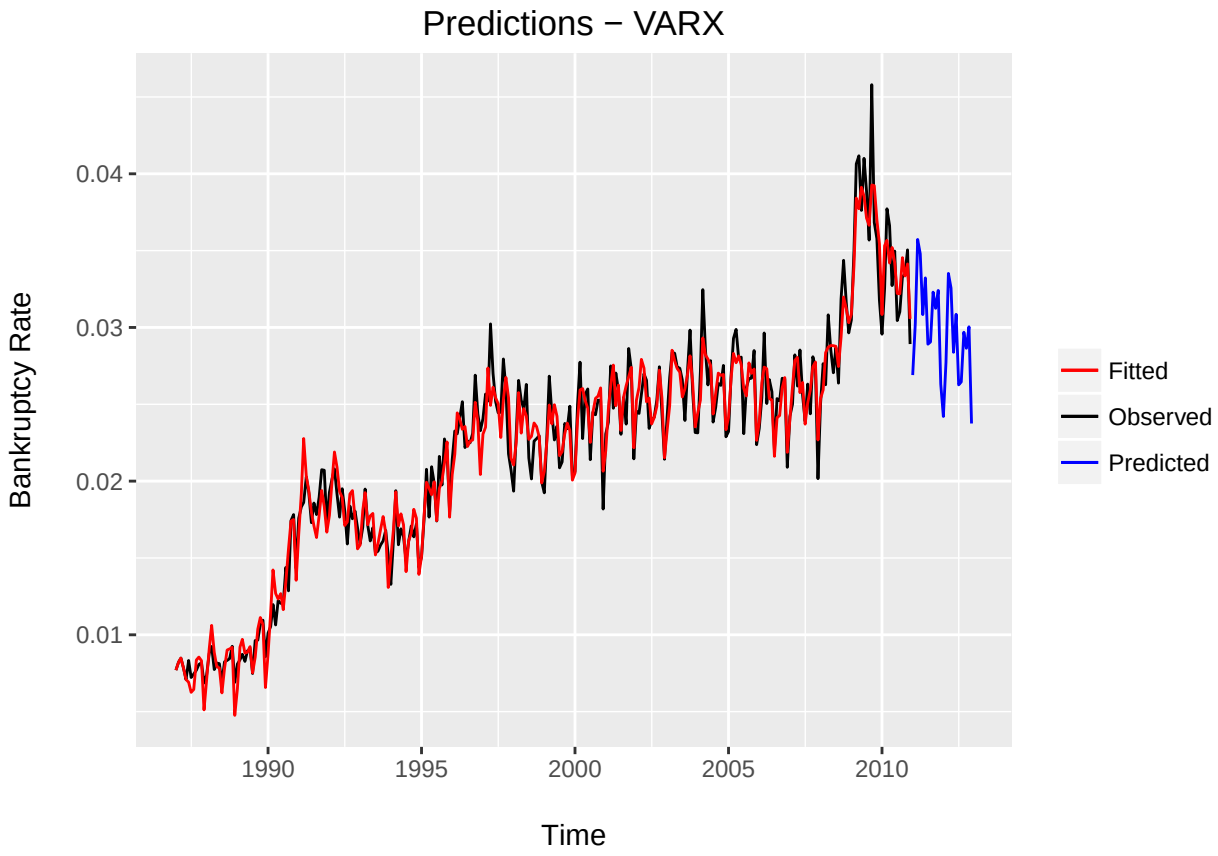
The best SARIMAX model is built on the best SARIMA model with the inclusion of external variables. We added to the model each of the 6 available combinations of external variables from the given data, selecting the combination with the best predictive accuracy. This model was the SARIMA model from above with the inclusion of population and unemployment as exogenous variables (i.e., variables which have an impact on the response but are not themselves influenced by it).



VARX

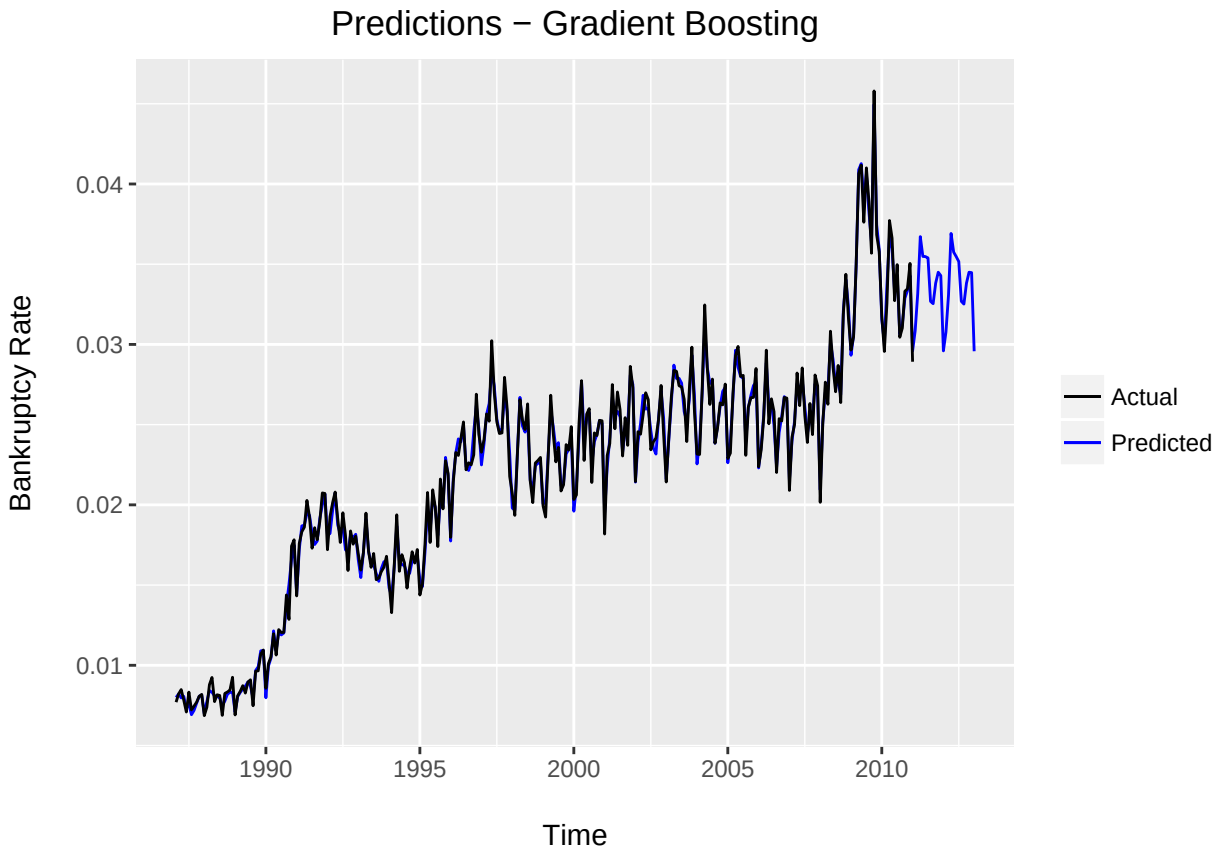
Variable auto-regression (VAR) is a modeling approach by which the variables which are thought to mutually impact one another are all modeled in terms of one another ("endogenous" variables). The "X" on the end indicates, as in the SARIMAX, the inclusion of completely external variables which are not modeled alongside the other endogenous variables but *are* used for prediction.

We elected to only treat the month as an exogenous variable and then to try different combinations of variables as endogenous entries to the model. The combination found to give the lowest MSPE was bankruptcy, population, housing price index, and unemployment rate.



Gradient Boosting

While not itself a time series model, we hypothesized that gradient-boosted decision trees would be appropriate for forecasting this particular time series data. The most notable problem with this approach to modeling most time series is that decision trees fail to accomodate trend in data (since the model may only make decisions based upon values which it has seen before). However, since bankruptcy rate is typically bounded (albeit increasing slowly over time), it may be worth attempting to fit such a model in this case. For background, this is an ensemble of decision trees where each new decision tree predicts on the residuals generated by the prior trees.



Model Selection

Since the primary foucs of this analysis is forecasting the next two years, we decided to limit ourselves the models with higher prediction accuracy (measured by the lower RMSE in the validation set). However, in order to avoid variation of our predictions just by relying on the "best" model with the lowest RMSE, we decided make the final prediction by taking the median predictions of the 3 models with the lowest RMSE. This gives us more stable predictions of the future.

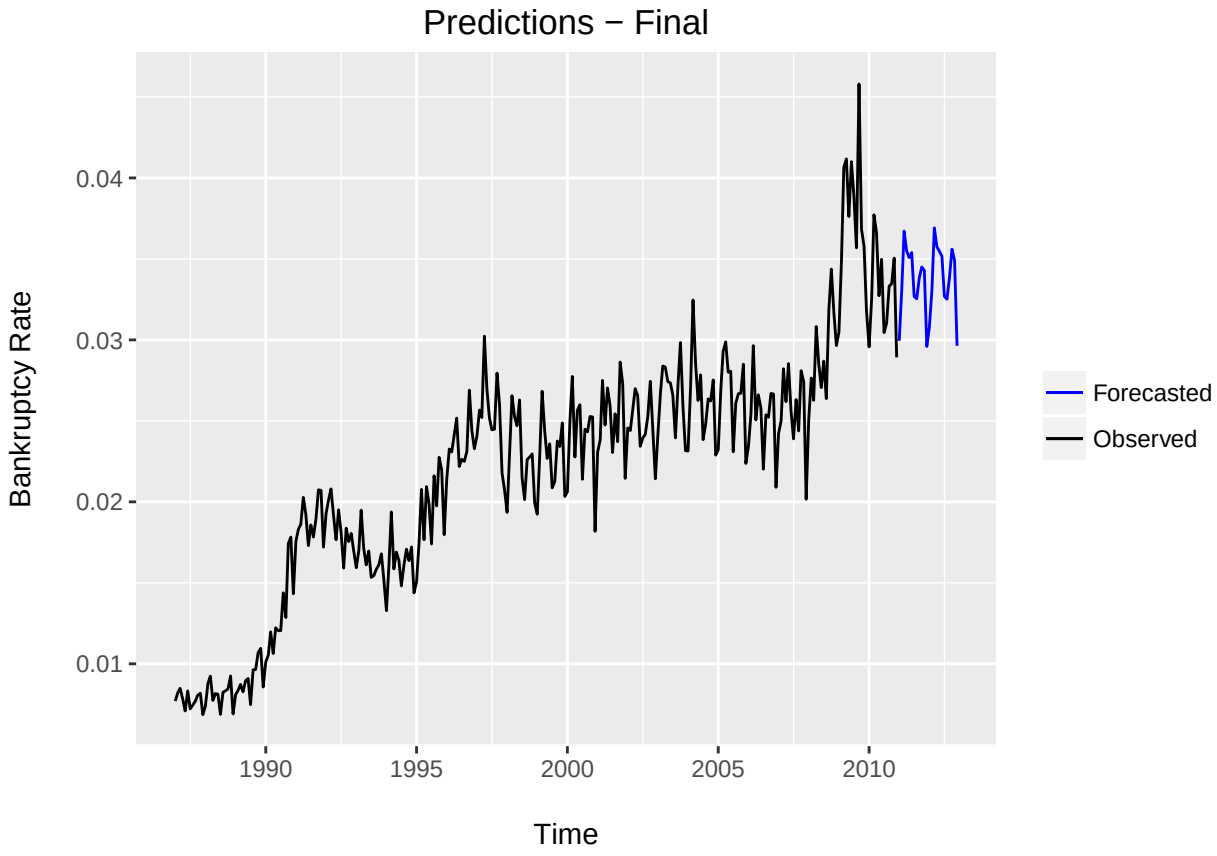
Table 1: Summary of Validation Errors

Model	RMSE
SARIMA	0.00161
SARIMAX	0.00169
VAR	0.00239
VARX	0.00275
Gradient Boosting	0.00132
Exponential Smoothing	0.00122

We do acknowledge that there is a particular drawback to our choice of final model. Since we have combined models, more or less "mixing and matching" our values together, we cannot construct prediction intervals for the final model. Were we to use only an exponential smoothing model, we could have produced valid prediction intervals, but our SARIMA model fails to have uncorrelated residuals and therefore cannot generate valid prediction intervals, and the gradient boosting model does not possess any system by which it can natively generate prediction intervals for time series data. So, our final model goes without prediction intervals.

Results

The plot below shows the actual values alongside our final predicted values for future observations.



The above plot visually verifies that this was a reasonable way to predict future observations of the Canadian bankruptcy rate since it maintains a similar level and seasonality to prior observations and thus our own expectations. After fitting a variety of models to the data, we belive this to be the most strongly predictive of future outcomes as an outlier-resistent combination of the highest-performing models we had produced.

Table 2: Final Predicted Values

Month	Prediction
2011-01-01	0.02994628
2011-02-01	0.03315142
2011-03-01	0.03671822
2011-04-01	0.03548503
2011-05-01	0.03508369
2011-06-01	0.03538984
2011-07-01	0.03268608
2011-08-01	0.03253987
2011-09-01	0.03382394
2011-10-01	0.03449935
2011-11-01	0.03428653
2011-12-01	0.02959827
2012-01-01	0.03077495
2012-02-01	0.03308281
2012-03-01	0.03691703
2012-04-01	0.03575245
2012-05-01	0.03547296
2012-06-01	0.03516519
2012-07-01	0.03268608
2012-08-01	0.03251579
2012-09-01	0.03387697
2012-10-01	0.03560268
2012-11-01	0.03490744
2012-12-01	0.02963245