Whitepaper
Hybrid Econometric Model

The hybrid model could be used to evaluate economic impact. Examples include infrastructure projects, special events and tourism.

By Sanjeev Naguleswaran and John Colombo

Table of Contents

1. Introduction

In this paper, we describe the use of Machine Learning to predict effects of economic policy. The application of Machine Learning in Econometrics is a relatively recent concept (Munnel, 96), (Ladd, 98), mainly due to the requirement for identification of causal conditions.

Machine learning, data mining and predictive analytics all use data to predict some variable as a function of other variables. The key aspects being:

- Insight, importance and patterns
- Inference - how a dependent variable changes as some independent variable changes, is sometimes possible.

Traditional Econometrics uses statistical methods for prediction, inference and causal modelling of economic relationships. Inference is the ultimate goal and in particular; causal inference is a goal for economic decision-making. Conventional statistical and econometric techniques such as regression often work well but there are issues unique to large datasets that may require different tools.

1.1 Summary of Terms

Data analysis in statistics and econometrics can be broken down into four categories: 1) prediction, 2) summarisation, 3) estimation, and 4) hypothesis testing.

Machine learning is concerned primarily with prediction. Machine Learning is often primarily concerned with developing high-performance systems that can provide useful predictions in the presence of challenging computational constraints.

Econometrics, statistics, and data mining are generally looking for insights that can be extracted from the data. Data mining is also concerned with summarisation, and particularly in finding interesting patterns in the data. Much of applied econometrics is concerned with detecting and summarising relationships in the data. The most common tool used to for summarisation is (linear) regression analysis. Machine learning offers a set of tools that can usefully summarise various sorts of nonlinear relationships in the data (Varian, 2014).

Data science, a somewhat newer term, is concerned with both prediction and summarisation, but also with data manipulation, visualisation, and other similar tasks.

1.2 Prediction

The goal of prediction is to get good out-of-sample predictions. It is easy to construct a predictor that works well in sample, but fails out-of-sample. To take a trivial example, n linearly independent regression equations will fit n observations perfectly but will usually have poor out-of-sample performance. This phenomenon is referred to as the “over fitting problem”.

Since simpler models tend to work better for out of sample forecasts, models with excessive complexity can be penalised and this process is known as "regularisation". Economists tend to prefer simpler models for the same reason, but have not been as explicit about quantifying complexity costs. It is also conventional to divide the data into separate sets for the purpose of training, testing and validation. Training data is used to estimate a model, the validation data to choose the model, and the testing data to evaluate how well the chosen model performs. (Often validation and testing sets are combined.)

If there is an explicit numeric measure of model complexity, it can be viewed as a parameter that can be "tuned" to produce the best out of sample predictions. The standard way to choose a good value for such a tuning parameter is to use k-fold cross validation.

1. Divide the data into k roughly equal subsets (folds) and label them by s = 1..k. Start with subset s = 1.
2. Pick a value for the tuning parameter.
3. Fit model using the k -1 subsets other than subset s.
4. Predict for subset s and measure the associated loss.
5. Stop if s = k, otherwise increment s by 1 and go to step 2.

Common choices for k are 10, 5, and the sample size minus 1. After cross validation, k values of the tuning parameter remain and the associated loss indicates an appropriate value for the tuning parameter (Varian, 2014).

1.3 Causality

Econometrics has developed several tools for causal inference such as instrumental variables, regression discontinuity, difference-in-differences and various forms of natural and designed experiments (Angrist and Krueger, 2001). Machine learning has, for the most part, dealt with pure prediction. Ironically, theoretical computer scientists, such as (Pearl 2009) have made previous significant contributions to causal modelling, albeit these theoretical advances have not as yet been incorporated into machine learning practice.

If data driven decision-making is the outcome required, usually the causal impact needs to be measured. (James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, 2013).

These predictive outcomes can be categorised as:

- Ceteris Paribus - Causal effect with other factors held constant, which is analogous to a partial derivative. This can be compared to explicit manipulation- If I explicitly change price, how do I expect quantity sold to change?
- Mutatis mutandis: Correlation effect with other factors changing, as they will analogous to a total derivative. This can be compared to Passive observation, asking the question "If I observe price change, how do I expect quantity sold to change?

Therefore, causality requires manipulation (Holland, 86).

Machine Learning although extremely effective at identifying correlations between variables requires adaptation for determination of causal relationships. For example, it could identify, through correlation, the fact that there are more police in areas of higher crime. However, this does not mean that police cause crime. It cannot also inform the policy decision as to whether more police should be added to a given district.

Another example is that there is correlation between people dying in hospitals does not mean that hospitals cause death, where the related policy decision is whether one should go to hospital for treatment.

1.4 Economic Example

Munnell et al. [1996] examined mortgage lending in Boston to see if race played a significant role in determining who was approved for a mortgage. The primary econometric technique was a logistic regression where race was included as one of the predictors. The coefficient on race showed a statistically significant negative impact on probability of getting a mortgage for black applicants. This finding prompted considerable subsequent debate and discussion (Ladd, 1998).

The rest of the paper will explore experimental methods to determine causality in an Econometrics scenario using Machine Learning.

2. Decision Trees

Decision tree based methods were selected for Machine Learning due to the fact they are invariant under scaling (i.e. more data inputs can be added over time) and transformations of inputs, is robust to inclusion of irrelevant features, and produces transparent models. For example, in building a model to evaluate significance of investment categories, data on multiple types of investment, private and government consumption and socio-economic variables need to be considered.

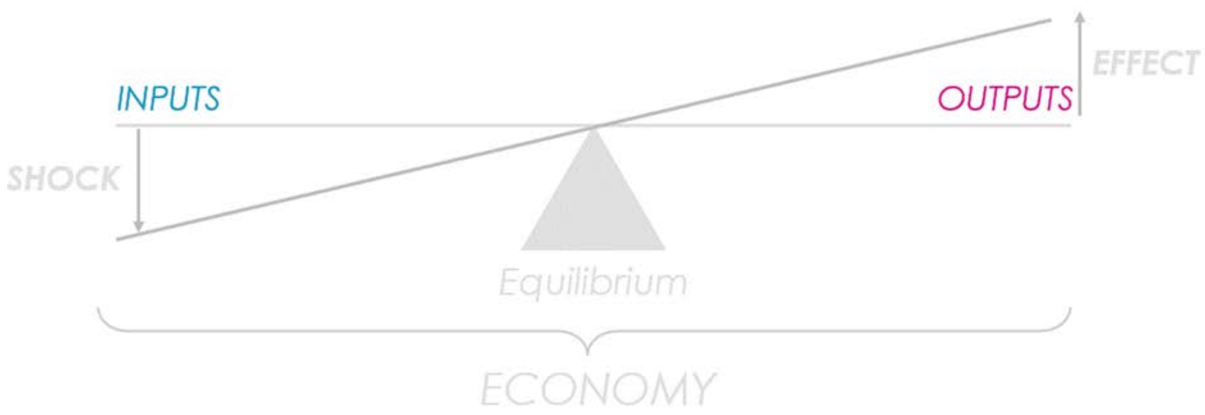
Techniques of varying complexity were evaluated in order to analyse the decision trees. At one extreme, there are simple methods that select a relative handful of the possible variables and build a model with fairly simple relationships between the input data and the forecast. These models run comparatively quickly, and the results are usually directly interpretable by a human. Most regression analyses and associated decision- tree methods, which generate a simple tree of "if-then" clauses, belong to this category.

We propose to use more complex Ensemble methods that take a number of simple models and combine them in some way to yield a final overall picture. Random Forests are a way of iteratively averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. Each iteration creates a simple decision tree on randomly selected subsets of input variables and input data. The final result is the aggregation of all such trees.

3. Defining the economic models

The modelling is based around a mathematical specification of key relationships within the economy (what determines levels of supply, demand etc.) to provide General Equilibrium. It is calibrated to real data to ensure that the model provides a good representation of the economy.

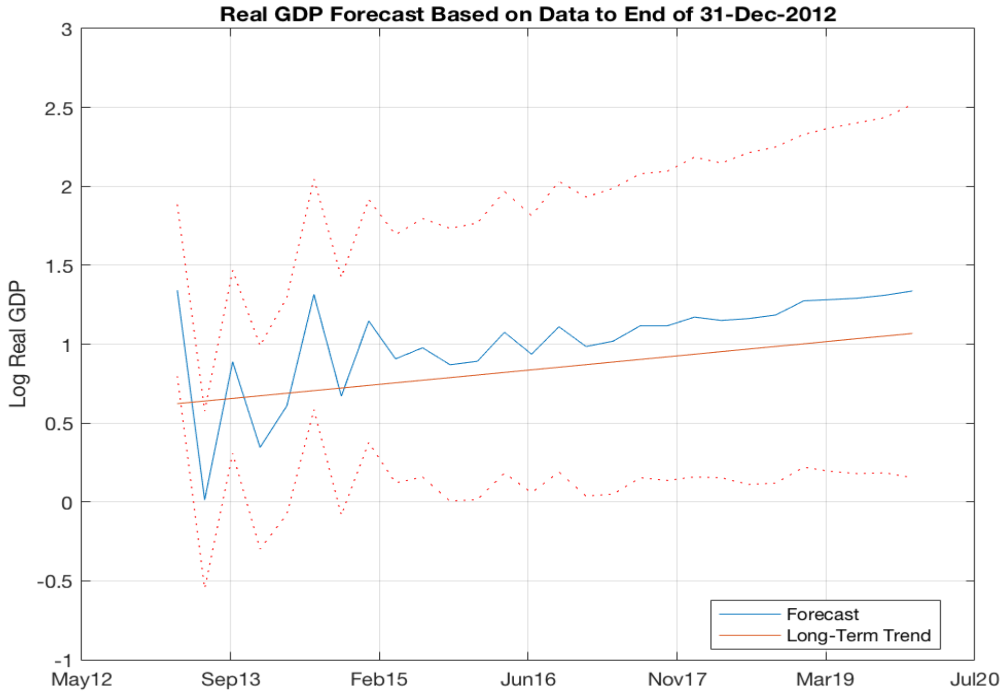
Variables



3.1 Hybrid Approach

We used a macro economic model to forecast the economy based on trend. The economic conditions (represented by GDP) resulting from external policy and investment are first estimated by solving a nonlinear system of equations. This model includes eight categories of time series: output (GDP), prices, wages, hours worked, interest rates, consumption, unemployment and investment (Del Negro, 2007).

The subsequent stage of the model traces out a specific time path of the economy following the change in the policy or introduction of a specific project. The economic adjustment can then be determined by the difference between the two alternative time paths (Jago 2006). We used a Machine Learning method based on regression trees to identify the impact on the baseline for investments in different industry categories. 'Ensemble' modelling methods to combine a number of decision trees (to provide the modelling foundation for investment 'what-if' scenarios) to yield a final overall picture of the effects of investment were then applied. The trees were aggregated iteratively where each iteration creates a simple decision tree on randomly selected subsets of input variables and input data (Breiman, 1996). This modelling recognised the behaviour of the budget sector as having relevance for the estimated economic impacts of an event and forecasts the outcomes. For example, when additional infrastructure spending is required.



3.3 Model Steps

1. Estimate significance (S) using Machine Learning and investment categories (Xi). N is the number of categories.
2. Equation based on S and X to reproduce the response variable (Y) - Y could be GDP or Jobs, for example.

$$Y = \sum_{i}^{N} S_i X_i$$

3. Create Two Groups randomly – Treatment (T) and Control (C).
4. Change each category of the T group, one at a time, by a fixed amount (Standard Deviation, $Million etc..)

	Treatment Group	Control Group
Pre	T_B	C_B
Post	T_A	C_A

After – Before = (TA – TB)
Treatment – Control= (TA – Ca)
Impact= (TA-TB)-(CA-CB)

5. Now write Y as

$$Y = Y_t + \sum_{i \neq t}^{N} S_i X_i$$

where

$$Y_t = \beta_t S_t X_t$$

$$\beta = \beta_0 + \beta_1 treat_t + \beta_2 post_t + \beta_3 treat_t * post_t$$

And
treat = 1 if treated, 0 if control,
post = 1 if after, 0 if pre
β_3 = treatment effect because

	Treatment	Control	Difference
Pre	$\beta_0 + \beta_1$	β_0	β_0
Post	$\beta_0+\beta_1+\beta_2+\beta_3$	$\beta_0+\beta_2$	$\beta_1+\beta_3$
Difference	$\beta_2+\beta_3$	β_2	β3

This set of linear equations can be solved using regression (Varian 2014).

3.4 Counterfactuals and rejection of Null Hypothesis

Important considerations in the determination of causal effects are counterfactuals and confounding variables. Counterfactuals are used in applied econometric research, particularly policy oriented research, to examine "what if" type questions. For example, in studying birth rates in a particular country, one may be interested in how birth rates would be affected had all females been above the poverty line, if they had high school educations, were all of a particular religion, etc. (Mincer Equation). One of the tasks involved in understanding causes is to compare the observed results to those you would expect if the intervention had not been implemented - this is known as the 'counterfactual'.

Many discussions of impact evaluation argue that it is essential to include a counterfactual. Some people however argue that in turbulent, complex situations, it can be impossible to develop an accurate estimate of what would have happened in the absence of an intervention, since this absence would have affected the situation in ways that cannot be predicted. In situations of rapid and unpredictable change, when it might not be possible to construct a credible counterfactual it might be possible to build a strong, empirical case that an intervention produced certain impacts, but not to be sure about what would have happened if the intervention had not been implemented.

For example, it might be possible to show that the development of community infrastructure for raising fish for consumption and sale was directly due to a local project, without being able to confidently state that this would not have happened in the absence of the project (perhaps through an alternative project being implemented by another organisation).

3.5 Experimental Identification of Counterfactuals

Counterfactuals can be identified by the following steps:

1. Developing a counterfactual using a control group.
2. Randomly assign participants to either receive the intervention or to be in a control group.

Control Group: a group created through random assignment that do not receive a treatment, or receive the usual treatment when a new version is being evaluated. This is an essential element of the Randomised Controlled Trial approach to impact evaluation.

The treatment effect can be hypothesised and the efficacy confirmed through acceptance or rejection of the null hypothesis. However, due to the importance of inference in Econometrics the Bayesian a-priori view should be considered in addition to the familiar frequentist view of statistics.

Both frequentist and Bayesian hypothesis testing incorporate an element of self-doubt, in that there is a requirement to show that there is some evidence that an alternative hypothesis is in some way a more plausible explanation for the observations than random chance. This is done in the frequentist methodology through using significance level. The Bayesian analogy is to use a scale of interpretation for the Bayes factor, such that a hypothesis won't be strongly suggested unless the Bayes factor over the null hypothesis were sufficiently high.

3.6 Model Uncertainty

An important insight from machine learning is that averaging over many small models tends to give better out-of-sample prediction than choosing a single model.

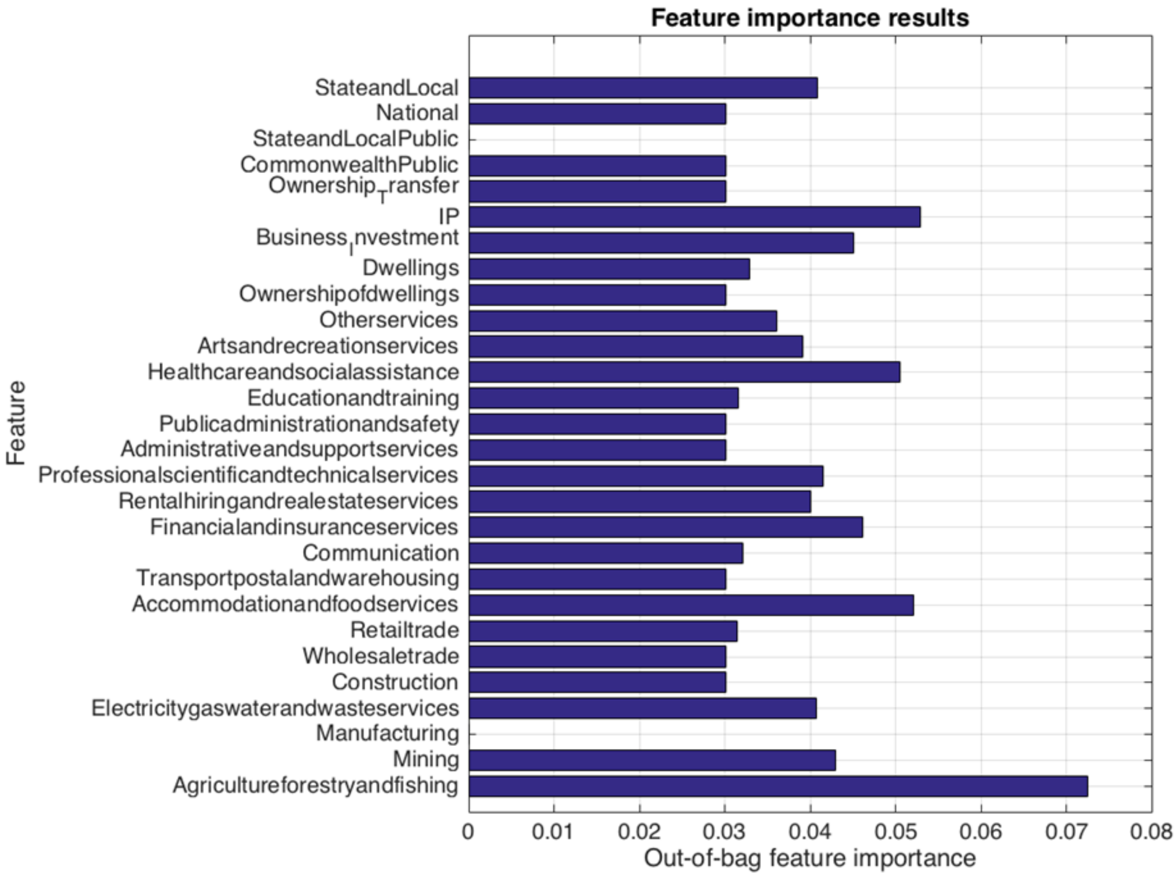
In 2006, Netflix offered a million dollar prize to researchers who could provide the largest improvement to their existing movie recommendation system. The winning submission involved a complex blending of 800 models" though they also point out that predictions of good quality can usually be obtained by combining a small number of judiciously chosen methods." (Feuerverger, 2012) It also turned out that a blend of the best and second-best submissions outperformed both of them.

Analogously, it was recognised many years ago that averages of macroeconomic model forecasts outperformed individual models, but somehow this idea was rarely exploited in traditional econometrics. The exception is the literature on Bayesian model averaging an overview of which is given in (Steel, 2011).

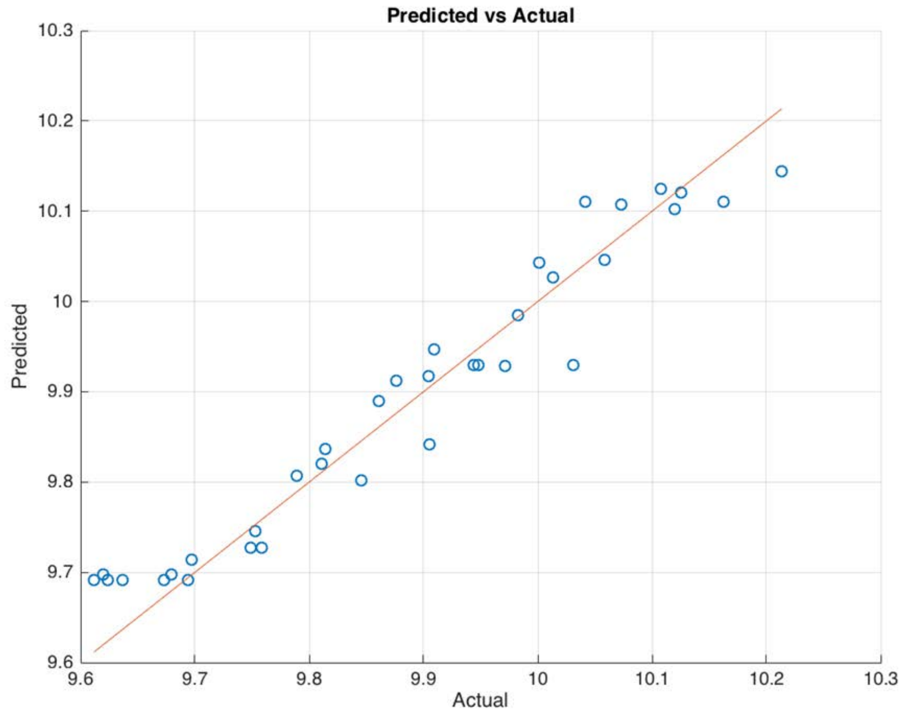
3.7 Model Outputs

3.7.1 Model Checking

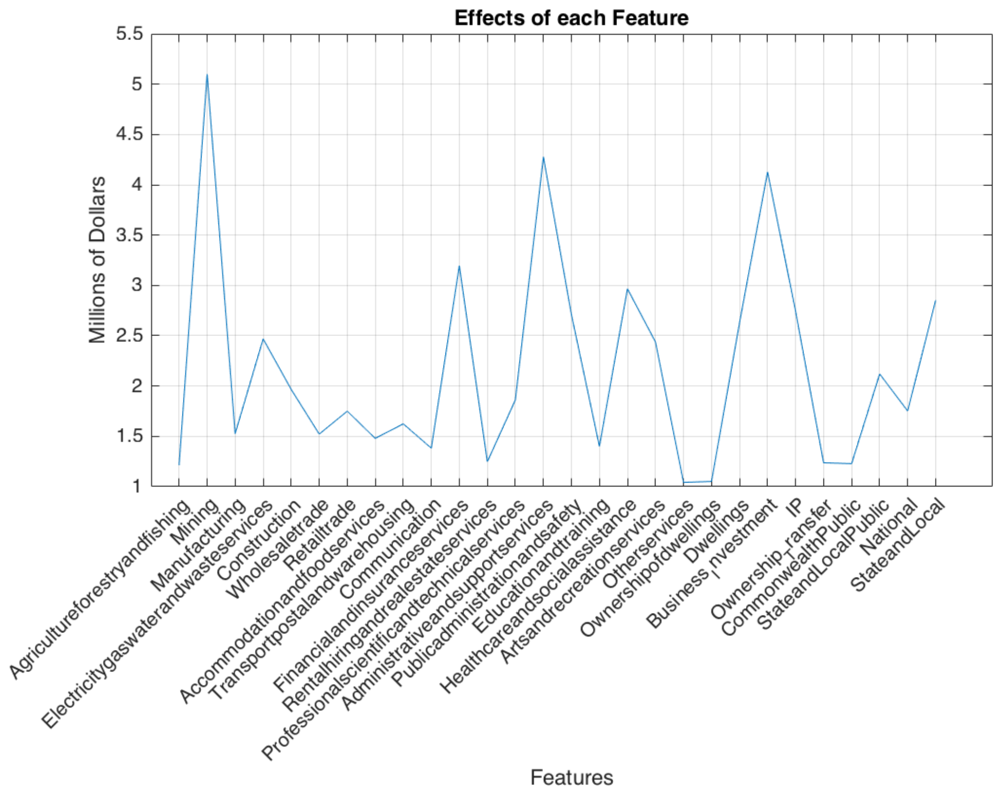
The relative significance of features calculated by the decision trees is given in the following figure:



The accuracy of the prediction is shown in the following figure:

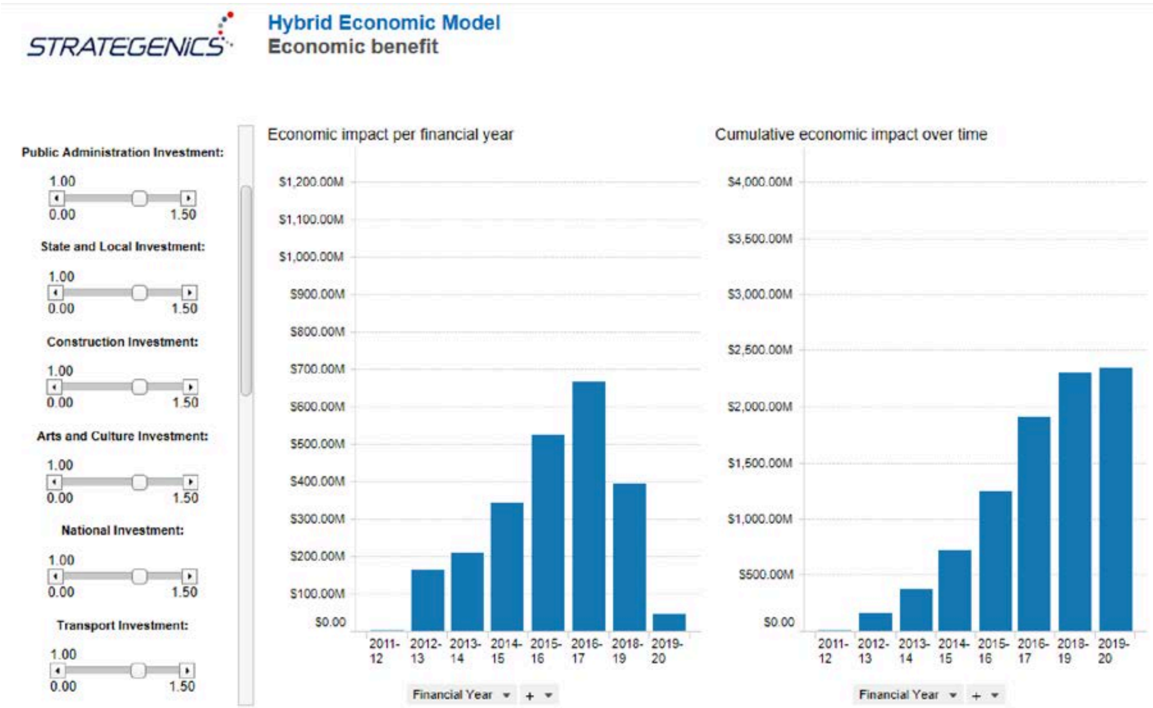


The monitory impact to the economy is shown in the following figure:



3.7.2 Example Output Charts

Interactive Spotfire output are given below:



4. Summary

The following insights were gained from the preliminary study:

- The data involved may require more powerful data manipulation tools for representation in a manner suitable for modelling.
- There may be potential predictors than appropriate for estimation, so some kind of variable selection would be required.
- Large datasets may allow for more flexible relationships. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning and so on may allow for more effective ways to model complex relationships.

The hybrid model could be used to evaluate economic impact. Examples include infrastructure projects, special events and tourism.

5. References

1. Angrist , Jand Krueger, A. Instrumental variables and the search for identification: From supply and demand to natural experiments.
2. Journal of Economic Perspectives, 15(4):69{85, 2001. URL http://www. aeaweb.org/articles.php?doi=10.1257/jep.15.4.69.
3. Angrist, J. and Pischke.J., Mostly Harmless Econometrics. Princeton University Press, 2009.
4. Breiman, L, Friedman, J. H.. Olshen, R. A and Stone. C. J. Classi cation and Regression Trees. Wadsworth and Brooks/Cole, Monterey, 1984.
5. Breiman, L, "Bagging Predictors," Machine Learning, Vol. 24, No. 2, August 1996, pp. 123–140.
6. Feuerverger A, Yu He, and Shashi Khatri. Statistical signi cance of the Net ix challenge. Statistical Science, 27(2):202{231, 2012. URL http: //arxiv.org/abs/1207.5649.
7. Hunn, C. & Mangan, J. (1999). 'Estimating the economic impact of tourism at the local, regional and state or territory level, including consideration of the multiplier effect' in: Valuing tourism: methods and techniques, Occasional paper No. 28, Bureau of Tourism Research, Canberra.
8. Holland, Paul W. [1986]. "Statistics and causal inference", Journal of the American StatisticalAssociation 81(396):945–960.
9. Munnell A.H, Tootell GMB, Browne, L, and McEneaney, J. Mortgage lending in Boston: Interpreting HDMA data. American Economic Review, pages 25{53, 1996.
10. Price Waterhouse (1989). 'Economic impact assessment of the 1988 Australian Formula 1 Grand Prix', A report
11. prepared for the Australian Formula I Grand Prix Board, Adelaide, Victoria.
12. Jago L and Dwyer L, Economic Evaluation of Special Events, A practitioners Guide, 2006
13. James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, 2013
14. Del Negro M, Schorfheide F, Smets F, and Wouters (2007), "On the Fit of New Keynesian Models," Journal of Business & Economic Statistics, Vol. 25, No. 2, pp. 123-162.
15. Ladd, H., Evidence on discrimination in mortgage lending. Journal of Economic Perspectives, 12(2):41{62, 1998.
16. Murphy K. Machine Learning A Probabilistic Perspective. MIT Press, 2012. URL http://www.cs.ubc.ca/~murphyk/MLbook/.
17. Pearl. J., Causality. Cambridge University Press, 2009a.
18. Pearl J,. Causal inference in statistics: An overview. Statistics Surveys, 4:96{146, 2009b.
19. Steel. M.F.J., Bayesian model averaging and forecasting. Bulletin of E.U. and U.S. In ation and Macroeconomic Analysis, 200:30{41, 2011. URL http://www2.warwick.ac.uk/fac/sci/statistics/staff/ academic-research/steel/steel_homepage/publ/bma_forecast.pdf.
20. Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28(2): 3-28.