Abstract—This article describes the results of a data mining project designed to explore the key drivers of the Australian Early Development Index (AEDI), a numerical indicator of early childhood development vulnerability. The work was conducted during GovHack 2013, a 48-hour Australian Open Data competition where participants were required to use published open data sets provided by various Australian government and other agencies.

We applied advanced machine learning techniques (random forests, generalised boosted regression models and multivariate adaptive regression splines) to the South Australian state and national data to gain insights into the key drivers of AEDI and to quantify the levers that the state government, community and individuals could apply to improve the situation.

We found that after accounting for the population specifics and socioeconomic conditions, for example unemployment level and Index of Relative Socioeconomic Disadvantage, the most important factors impacting early childhood development were lack of motor vehicle in the household, inability to afford buying medication and maternal smoking during pregnancy. We quantified the impact of each of these factors and suggested relevant potential Government interventions.

We then visualised our findings and created a Web app that allowed various intervention strategies to be interactively explored, based on the derived relationship between early child development index and its key drivers.

Keywords-random forests, generalised boosted models, multivariate adaptive regression splines, R, Tableau, early childhood development, AEDI, open data, GovHack, social health atlas.

I. INTRODUCTION
As part of a broader Open Government and Open Data philosophy, a National Open Data competition was held in June 2013 under the title GovHack [1]. Australian Federal and State government agencies had published suitably licensed and Open data sets to a common repository and competition participants drawn from industry, government, academia and the community were required to use these in whatever way they wished. The competitors formed teams and, under a strict time limit of 48 hours, aimed to deliver a working prototype or concept for the use of these data. In 2013 the event ran concurrently in eight cities and attracted over 1000 participants.

We elected to use data science and data visualisation techniques to explore the social issue of early childhood development.

The paper first provides information on the project background and objectives, then outlines data sources used, data analysis performed and actionable insights delivered. Then visualisation of the solution is described. We then provide the list of software tools that we used for data pre-processing, analysis and visualisation.

II. BACKGROUND
A. Acknowledge the problem
In 2003, the South Australian government embarked on a bold idea: a Thinkers In Residence program [2] that set out to bring new ideas into the state and to translate them into practical solutions to improve the lives of citizens. Individual experts were invited to participate in three month residencies that aimed to generate new thinking and provoke change. During his residency in 2006–07 [3], Fraser Mustard, a leading expert in early childhood development, noted:

The challenge for all societies is to close the gap between what we know about the determinants of early child development and what we do.

Early childhood development became a strategic priority for the South Australian government. It has been established that the first five years of a child's life has major influence on their later success in life [4]–[7]. Among the consequences of the Mustard residency was the creation of the Australian Early Development Index (AEDI) and its implementation across South Australia in 2009.
B. Measurement and monitoring

AEDI was devised as a means of quantifying early childhood development outcomes nationwide. AEDI provides information about how local children have developed by the time they start school across five areas of early childhood development: physical health and well-being, social competence, emotional maturity, language and cognitive skills, communication skills and general knowledge. The AEDI results report on the number of children scoring in the following percentile ranges: 0 to 10th percentile (developmentally vulnerable), 11th to 25th percentile (developmentally at risk), 26th to 50th (on track lower range) and above the 50th percentile (on track higher range).

AEDI data are collected every three years and commenced in 2009. With the recent addition of 2012 data, policy makers have now begun to make broad comparisons across time based on counts (or proportions) of children who are deemed to be developmentally vulnerable [8], hoping to demonstrate an overall decrease in childhood developmental vulnerability.

C. Determination of key drivers of AEDI

Such monitoring is a necessary first step but to date there has been no attempt to quantify the drivers of the magnitude of AEDI. The consensus view is that poor AEDI in a particular geography is correlated with low socioeconomic status, high unemployment and associated social indicators, i.e. large and complex social issues that are not easily solved. This project aimed to quantify the key drivers and to evaluate their suitability for the creation of targeted intervention programmes.

D. Economic impact and practical intervention

Based on AEDI data published in the Australian Social Health Atlas, as many as 23% of South Australian children in their first year of school are developmentally vulnerable on one or more domains (social, emotional, physical, language/cognitive and communication/general knowledge). A rough estimate of the cost incurred by this to the State Government is up to $20m per year. Given the above, there was a clear need to understand what factors influence the risk of early childhood developmental issues, find the levers that can be used in a cost-effective and efficient way to improve the situation, quantify the effect of each of the levers or their combinations, and recommend to the Government both actionable and defensible intervention steps.

III. APPROACH

In contrast with using classical statistical methods such as factor analysis and linear modelling, or investigating correlations between the outcome and other variables, or applying causal modelling, our chosen approach was based on predictive data mining. Predictive data mining relies on analysing historical data to gain insight into what influences the outcome of interest and how it can be improved. Typically such a model uses vast numbers of variables and observations to identify the drivers of the outcome of interest out of many potential factors and establish to what extent the change of the drivers and their interactions can influence the outcome.

Such techniques have been successfully used in the last 20 years to guide decision making across the industries, for example in actuarial science [9], marketing [10], insurance [11], finance [12], forensic accounting [13]. Our analysis was implemented using such powerful, proven and robust predictive data mining techniques: random forests, generalised boosted regression and multivariate adaptive regression splines (MARS). These methods, in particular random forests and generalised boosted regression are known to be minimally affected by multicollinearity and noise in the data as opposed to linear methods. To further enhance the robustness and validity of the results in the situation of extreme time constraints, we applied all three methods and compared the results to check for any inconsistencies.

IV. ANALYTICAL SOLUTION DEVELOPMENT

We started from identification of the key data sources in the public domain that contained factors that could potentially influence early childhood development. We then retrieved, combined, cleaned and pre-processed the relevant data, removed redundant information i.e. the variables which were highly correlated to each other and applied three powerful machine learning methods to identify the most important drivers of the early development vulnerability in South Australia. Then for the ease of explanation of results to the users we expressed the results as a linear regression with main effects built using the key predictors that were identified at the previous step.

All analysis was done at the Statistical Local Area (SLA) level, because the data on AEDI was provided at that level.

A. Data source identification and retrieval

We used the following open data sources provided by Federal and State Government:

1) Social Health Atlas of Australia [14]
2) Australian Early Development Index Data [15]
3) Health service locations [16]
4) Geographic boundary files for statistical local areas (SLAs) [17]

Analysis data were principally sourced from the Social Health Atlas of Australia published data sets [14]. Grouped based on geography, the Atlas comprises comprehensive population statistics along several health-related and demographic domains, among them: age distribution, ethnicity, education, early childhood development, families, housing/transport, chronic disease, health risk factors **inter alia**.

B. Data analysis

All data analysis, including data retrieval, pre-processing and modelling was performed in R.

1) Data preparation and pre-processing: There were 2000+ variables in the Social Health Atlas data set. The data were stored as an Excel file that included 46 worksheets, with a separate worksheet per topic (for example Families, Education, Housing_Transport, Mothers_babies, Aged_care). We
consulted subject matter experts to exclude from consideration data in the Social Atlas that was unlikely to be related to our outcome (for example, data about aged care services). We then imported the remaining data into R and merged it together using Statistical Local Area (SLA) as the id variable. We checked the data for any quality issues, for example duplicates, obviously incorrect values, unusual distributions or missing values. The only issues we found were related to missing data values (about 2% of the values overall). We addressed this by imputing the missing values using the k-nearest neighbours method implemented in R in the imputation package.

2) Variable selection: An important reason for choosing random forests, generalised boosted regression and MARS as our modelling techniques was the fact that in the situation of having to select the important predictors out of myriad potential variables typically available in organisational data, these techniques can be successfully applied for variable selection (as discussed, for example, in [18]–[21]). To further enhance the robustness of variable selection, we performed it using each of the three methods and then compared the results to check for any inconsistencies.

3) Correlation analysis to establish redundant data fields: We applied correlation analysis to the data to identify the redundant information, i.e. the variables which were very highly correlated to each other and so could be removed from the analysis [22]. This was implemented using caret package in R as follows. The absolute values of pair-wise correlations were considered. If two variables had a correlation of 95% or higher, then the method reviewed the mean absolute correlation of each variable and removed the variable with the largest mean absolute correlation. This step reduced the dimensionality to 45 attributes.

4) Modelling the level of childhood vulnerability and identifying the most important predictors: To accurately and quickly identify the most important predictors out of the remaining variables we applied three methods that are known for excellent performance in predictor importance ranking:

- Random forests, an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees [21], [23], [24];
- Stochastic gradient boosting, a computational approach that has emerged as one of the most powerful methods for predictive data analytics [25], [26]; and
- Multivariate adaptive regression splines (MARS), a non-parametric regression technique that can be seen as an extension of linear models that automatically models non-linearities and interactions between variables [27], [28].

All modelling was implemented using R packages randomForest, gbm and earth.

To maximise the accuracy of the key predictor selection, we compared the output of each of the models and found that the predictors nominated as important were consistent across the models with the few top predictors standing out as the most important.

A major constraint of the project was that it had to be conducted within 48 hours. We therefore made several pragmatic decisions, among them, how best we could translate the R-generated modelling results into a user-friendly format that conveyed the salient points of the analysis in the way that was easy to understand for non-technical users. We took the following approach:

1) We elected to express the relationship between AEDI and the key drivers as a main-effects-only linear equation, i.e. no interaction terms (to ensure the validity of the model, the relevant assumptions checks were performed and other best practices for regression modelling were observed).
2) This form of equation was easy to interpret and explain and quick to implement in the visualisation tool set.

Alternative approaches, for example, expressing the relationship in the form of a decision tree or using more complex modelling equations, were rejected because of the severe time constraints. In the real-world commercial situation removal of such constraints would allow more elaborate (and accurate) model results representation.

The key levers of the outcome identified by our analysis that could be used for time- and cost-efficient interventions were:

1) maternal smoking during pregnancy
2) inadequate access to transport
3) inadequate access to medication.

V. RECOMMENDATIONS FOR INTERVENTION STRATEGIES

Based on the key levers of the outcome identified by our analysis we recommended the following intervention strategies:

- Community and nurses to deliver anti-smoking education to women of child-bearing age
- Provide taxi vouchers or mini buses to health centres on a regular basis, particularly for pregnant women and women with children of early age
- Provide subsidies for medication specific to pregnant women, mothers and young children.

VI. VISUALISATION OF THE SOLUTION

A. Geographic data preparation

Geographic data were published in SHP format, commonly used in commercial Geographic Information System (GIS) applications but not compatible with our visualisation software. Geographic polygon data were therefore translated to the more open comma-separated-value (csv) format using the shapetotab utility [29]. Superfluous geographic roles were purged from the resulting file by using the -simplify option. This reduced the size of the csv file and thus improved performance.

Geographic data for Statistical Local Areas (SLAs), a common geographical unit used by Australian government agencies to publish population census and other data, were translated in this way to produce a single csv file, each row of which comprised five variables: the SLA id value, latitude
and longitude of the polygon boundary coordinates, polygon point order and polygon number.

B. Geographic visualisation

We used Tableau software [30] to visualise the AEDI value per statistical local area. Tableau Desktop Professional 8 was used to build a Tableau workbook that contained the geographic visualisation; Tableau Server 8 was used to publish the workbook to users. Users connected to the workbook using a modern Web browser (for example, Google Chrome, Firefox 11 or later).

Two data sources were used for the visualisation, joined by the common key of SLA id value: AEDI data where each observation represented a single SLA; SLA polygon boundary coordinate data in csv format. We created a map-type visualisation by specifying polygon mark type for the polygon coordinate fields.

1) User interaction: In a business climate of restricted resources, policy makers are interested in devoting resources to strategies that yield the “biggest bang per buck”. This visualisation approach allowed exactly for this kind of scenario exploration and testing.

We aimed to not only visualise the magnitude of observed AEDI per geography, but also allow the user to modify the loadings associated with the equation relating AEDI magnitude to its key drivers and thus visualise the modelled AEDI value.

We used the main effects linear regression equation derived in R to calculate the modelled AEDI response value. We further added a delta to each main effect parameter value for the purpose of scenario testing. Default delta values were set at 1.0 (i.e. no modification to the models actual parameter values).

We allowed users to modify delta values by creating quick filter slider objects connected to the delta parameters in the Tableau visualisation interface. Slider calibration was set between 0 and 1, where 1 indicated no change to the equations parameter values and 0.5 (say) represented a reduction in parameter loadings by 50%.

2) Scenario testing: The final visualisation (Figure 1) showed quick filter slider objects corresponding to the delta parameter values attached to each main effect parameter of the AEDI magnitude: Maternal smoking during pregnancy, access to transport and access to affordable medication. Users were able to explore the impact of various hypothetical intervention scenarios and observe the resulting change in AEDI value for each geography. For example, to simulate the effect of halving the incidence of smoking among expectant mothers, the smoking quick filter slider could be set to a value of 0.5. Modifications could be made to individual drivers, or all three, to explore the impact of single or multiple intervention strategies.

VII. SOFTWARE USED

- R v2.15.3
- R packages including car, caret, earth, ellipse, gbm, ggmap, ggplot2, Hmisc, imputation, maptools, plyr, randomForest, rattle, rgdal, RODBC, sp, stringr, taRifx, XLConnect
- Tableau Desktop Professional 8
- Tableau Server 8
- Microsoft Excel 2013

REFERENCES


