Classification of Review Sentiments

Vivek Kumar
Civil and Environmental Engineering
Princeton University
vivekk@princeton.edu

Victor Charpentier
Civil and Environmental Engineering
Princeton University
vc6@princeton.edu

Abstract

Applying machine learning techniques for social good is one of the bright outcomes of the technological revolution of the past years. In homework 2 we have performed a prediction and analysis of 6 key outcomes at age 15 of the fragile family challenge. The methods implemented depend on the type of outcome. For each type of outcome over 6 prediction methods are tested. The performance of regressions and classifiers is assessed with xx and receiver operating curves. Overall, the random forest regression methods has the best results for continuous variables and XX has the best result for binary outcomes. Finally, the confidence intervals of the predictors are computed to quantify the quality of the prediction.

1 Introduction

The fragile family challenge (FFC) is a Princeton University-led initiative opening the data collected in the long term Fragile Family and Child Wellbeing Study (FFCWS) to a variety of data scientists. The study follows an ensemble of nearly 4700 families with children born in 20 cities across the U.S. between 1998 and 2000. The study has an over-representation of non-marital birth compared to national average. The goal of the study is to provide information on the family factors that influence the life of the child [5]. The FFC was officially closed in May 2017, and its winner announced in the fall of 2017. With an increased understanding of the data gained from the challenge, we aim to apply regressions and classifications to the dataset of the FFC to predict the 6 key outcomes. Regressions and classification constitute the backbone of supervised learning. However beyond the purely mathematical aspect of the challenge, feature selection and dataset preparation are some of the most important aspects of a real machine learning project. So important that some describe the refinement of feature or feature engineering as a "*black art*" [3]. With respect to this mystical reference we have worked to prepare the training data for prediction. The continuous outcomes (GPA, grit and material hardship) have been predicted using Random Forest Regressor (RF), Ada Boost Regressor (AB), lasso regression (LRCV), ridge regression (RRCV), elastic net (EN), extra trees regressor (ET) and multi-layer perceptron regressor (MLP). The binary outcomes (eviction, job training and layoff) were predicted with quadratic discriminant analysis (QLA), logistic regression (LR), random forest classifier (RFC), ada boost classifier (AB), multi-layer perceptron classifier (MLP) and extra trees classifier (EF).

2 Description of the Data

In the domains that generate a lot of data, social sciences are maybe one of the least explored by data scientist. Social sciences excel in causal inference, they have handled data from observations for a long time. Data science on the other end has experience with large dataset but is not focused so much on causal inference [4]. The hope of applying machine learning techniques to the FFCWS dataset is to combine both sciences for the understanding of fragile families outcomes. Improving the methods to treat large datasets is vital for social sciences and the potential outcomes could be transformative for society.

Inequalities in the U.S. are threatening the model of the American Dream. Some voices from the highest economic international agencies are calling for a new American Dream, "*based on equality and sustainable growth*" [1]. In this context, understanding what factors positively influence children growing up in some of the most difficult family environments may be key to creating policies more able to grant equal opportunities for all than currently.

3 Overview of the methods and fit to reference data

4 Detailed presentation of the regression forest

The random forest predictor used for regression, also called regression forest (RF), is based on the stand alone model of decision trees. Decision trees are prone to over fitting and sensitive to missing data. Therefore, a forest stems (pun) from the aggregate of many weak models to produce a better overall prediction. As opposed to a classification forest, a regression forest provides prediction of continuous variables. The input data seen on Figure 1 is continuous. Therefore the leaf nodes predict real values (as opposed to classes). The data-set is split based on homogeneity of data (with the standard deviation). This leads to subsets of the data contains similar values of the data to predict (cf. Figure 1(a)). The similarity is quantified by entropy, a measure of predictability. The form of the predictor can vary once the tree is fully grown. Several types of predictors can be used [2]. They can be constant, linear, polynomial or probabilistic-linear among others 2.

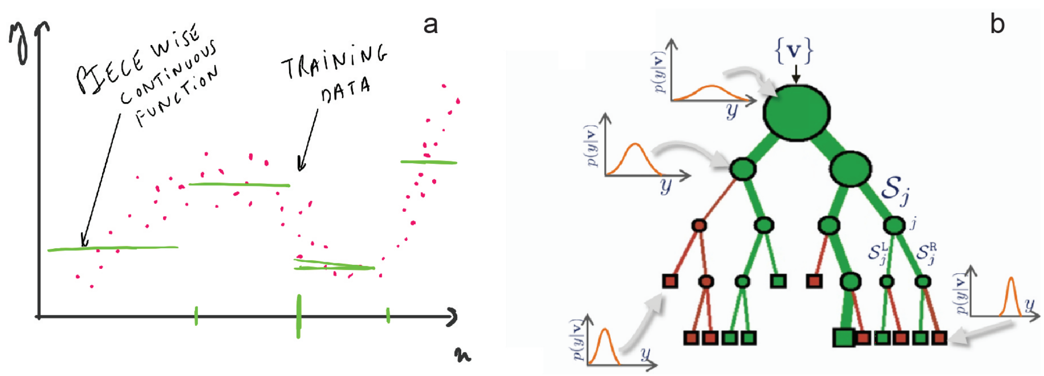


Figure 1: Detail of a regression forest - (a) the continuous data is approximated by a piecewise continuous function. The subdivision of the x interval stops when the entropy or SSE has reached a threshold value - (b) an example of regression tree from [2], the leaves of the tree determine values of continuous outputs for subdomains of x

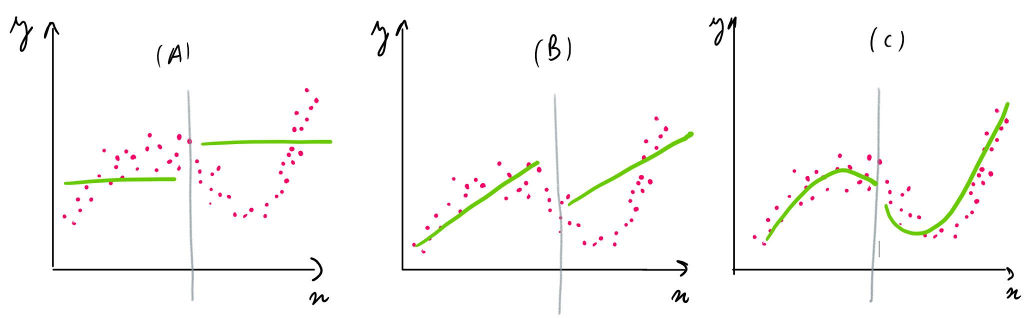


Figure 2: Example of three predictor models (a) constant - (b) linear and (c) polynomial

In the case of the constant predictor, the value of the predictor is given by the minimization of the sum of squared error (SSE) as given by Equation 1 for the subdomain D_k.

$$\hat{y}_k = \arg\min_y \sum_{i \in D_k} (y - y_i)^2 \tag{1}$$

Two parameters appears as main control nobs of the method: the number of regression trees in the ensemble and the depth of those trees. Each of those parameter has a distinct effect on the prediction. The depth of the trees controls the closeness of the fit. A ensemble of trees of depth 1 will correspond to a linear regression since each tree corresponds to its root node. In comparison, an ensemble of very deep trees will risk being overfitted. The number of trees in the forest influences the smoothness of the prediction, the more trees the smoother the direction of the prediction. Finally regression forest is a preferred method due to its speed, efficiency, ability to cope with missing data and flexibility of use.

5 Presentation of results

1. Build a model for predicting 1-type of variable well.
2. GPA, Grit, Material Hardship
   (a) Hypothesis : Positive Environment + Lack of Negative leads to positive GPA + ..
   (b) Hypothesis : Positive Environment + Plus some negativity + ..
   (c) Hypothesis : Negative Environment + Lack of parent + ..
3. Drop NA values for training: only in case of NA in all.

6 Results and Discussion

The first result we present is that of cross-validation performed on the split data to determine the best for each prediction. The total train data (obtained from `train.csv`) was divided randomly into a 50:50 split. The cross-validation score chosen for comparison was mean-squared-error (MSE). The results for each are displayed in table[]

Method	GPA MSE	Grit MSE	Material Hardship MSE
RandomForestRegressor	0.221031	0.153470	0.015116
AdaBoostRegressor	0.250311	0.156094	0.017874
LassoLarsCV	0.232776		
ElasticNet	0.379290		
ExtraTreesRegressor	0.219626	0.162074	0.016035
MLPRegressor	157140	190226	305336

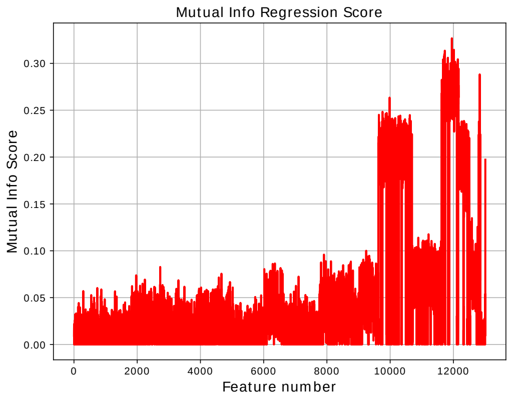
Table 1: Mean-Squared-Error score for various methods obtained in cross-validation

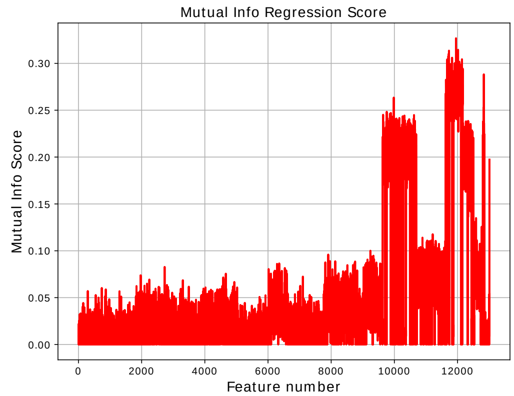
Using the best method the leaderboard scores were :

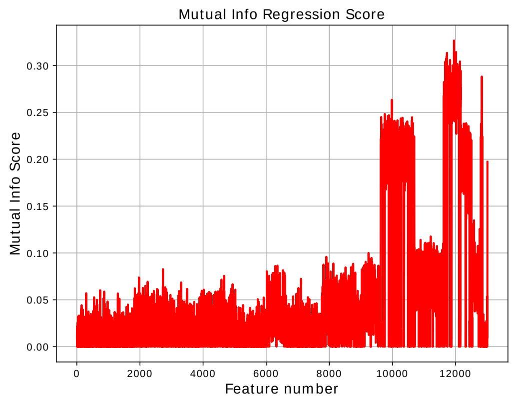
$$\text{GPA Score} \;: \qquad \text{Grit Score} \;: \qquad \text{Material Hardship Score} \;:$$

The bootstrapping was performed on the training data set to determine the minimum and maximum error obtained for each prediction. Here we have used RandomForestRegressor as the method for predicting the results. The results are plotted in the Figure [].

Further to understand, which features were most important the `mutual_info_regression` from `feature_seclection` of sklearn was used. The plot shows the mutual information between each feature and the target prediction.







Method	None	Mutual Info (k-best)	PCA(99% variance)
RandomForestRegressor		0.226496	0.015116
AdaBoostRegressor		0.240569	0.017874
LassoLarsCV		0.233773	
ElasticNet		0.245301	
ExtraTreesRegressor		0.223951	0.016035
MLPRegressor		51837	305336

Table 2: Mean-Squared-Error score for GPA with and without dimension reduction/ feature-selection

We also employed the principal component analysis for dimension-reduction. Instead of showing the individual results we compare how the methods performed when various feature selection or dimension reduction techniques were employed.

7 Conclusions

Based on the results presented above, it can be concluded that RandomForest is the best regression technique for the continuous variables. Feature selection helps but the reduction in the MSE from the cross-validation suggests not such a great increase in accuracy. However, Performing Principal Component Analysis definitely helps in reduction in computation time. As the data-set becomes larger it would be of paramount importance to perform such dimensionality reductions.

8 Acknowledgment

We would like to acknowledge the help provided by the COS 424 staff during the completion of this project both through Piazza and through office hours.

9 Bibliography

References

[1] How the american dream turned into greed and inequality. online. Accessed: 2018-04-03.

[2] A Criminisi, J Shotton, and E Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research*, 2011.

[3] Pedro Domingos. A few useful things to know about machine learning. *Communication of the ACM*, 55(10):78–87, 2012.

[4] Justin Grimmer. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1):80–83, 2015.

[5] Nancy E Reichman, Julien O Teitler, Irwin Garfinkel, and Sara S McLanahan. Fragile families: Sample and design. *Children and Youth Services Review*, 23(4-5):303–326, 2001.