PREDICT THE SIZE OF THE POPULATION OF HIGH EARNINGS WITH CORRELATED FEATURES

Zihan Ding, Lingjun Liu, Meiru Zhang, Mingrui Zhang, Yan Ran, Jianhong Wang
Imperial College London

ABSTRACT

As the increasing impact of behaviour of society, socioeconomic status becomes more important to be analyzed. In this project, we firstly analyze the correlations related to the size of the population of high earnings. Additionally, we use the captured features above to conduct the estimation of the size of the population of high earnings. In this correlation analysis, we find out education, poverty and income will greatly impact the population of high earnings. Due to the decent features we find, the estimation performance is good in our validation set.

1 INTRODUCTION

Socioeconomic status is increasingly significant and desired to be investigated, since we can utilize the obtained results to predict analyze the behaviors of the society, e.g. health, income and education. Government can therefore take into consideration of socioeconomic status to change the policy so that the society can be more stable and level of happiness can be higher. In this project, we aim to estimate the size of the population of high earnings in each district. To fulfill this objective, we firstly analyze the correlations between other features and the target feature that we would like to predict. Around two hundred features are selected based on the magnitude of correlations. Then modern dimension reduction approaches are applied. Lastly, we attempt multiple models to predict the size of the high earning population by the features that we select.

2 RELATED WORK

As inducted by Solon (2002), the earnings should be highly related to education and occupation, which is summarized by a survey in 1957. Another report shows that the earnings should be correlated to personality and behaviours (Bowles et al., 2001). Specifically, employers should increase the incomes according to the preference of employees, including family and household. Moreover, Featherman & Hauser (1976) indicates the inequality between male and female which may cause the difference on education, so as the earnings. The provided dataset of our task including the survey from American citizens who live in California. This dataset is split into 20 subsets where each subset candidates a category of features, e.g. race, health insurance and migration. We will compare the categories of the selected features with previous research results, and to investigate similarities and variations.

3 METHODS

In this section, we will briefly introduce how we select features and how we predict the size of the population of high earnings.

3.1 SELECT FEATURES

To select features, we repeatedly calculate the Pearson correlation coefficient between the feature of the size of the population of high earnings and other features. To acquire more variant results and features, we select three levels of thresholds, e.g. 0.5, 0.6, and 0.7. The feature with the coefficient higher than the threshold will be remained, otherwise it will be dropped. Noticed that

before computing the correlations, we separate the whole dataset into two partitions with 80 percent of training data and 20 percent of validation data. To avoid cheating, we only select features based on the training data. When selecting the features we also build up a dictionary to indicate which features should be kept in testing data. About missing data, before we do the correlation analysis, we drop out all of columns with null value.

3.2 PREDICTION

For prediction, because this is a regression problem, we select support vector regression (SVR) Drucker et al. (1997), linear regression (LR) Neter et al. (1989) and Bayesian linear regression (BLR) Raftery et al. (1997). The reason for attempting SVR is that there are many outliers existing in the dataset, and SVR can well solve out this problem without losing too much data. Linear regression is a classical model which can work well in many cases with good feature selections with high dimensions. Since the dimension of our features is quite high, we believe that the linear regression model can have a good performance. Compared with linear regression model, Bayesian linear regression model averages the results of linear regression models with different possible parameters, with a prior constraint on parameters search space. By this method, it can mitigate the effect on noise existing in real data, such that the distribution of training data may be biased compared with testing data. To evaluate the results, we use the mean-squared-error, which is a frequent metric used in assessing the regression performance.

4 ANALYSIS

Based on the correlation check between male population 16 years and over with earning over $ 100000 and all the other types of features, the features results in correlation over 0.5, 0.6 and 0.7 are selected to train the model. The common trend is that the higher the correlation selected, the less features being used for training model.

4.1 CORRELATION OF 0.7

When selecting the interesting features with correlation greater than 0.7, 25 out of 39 features are relevant to education. In more details, these features contains the sex by education, sex by age by field of study. The sex by education contains both male and female which means that the earning of male is also affected by the education level of females. The sex by age by field of study can reflect the occupation. This matches with the research by Solon in 2002 that the earnings are highly related to education and occupation. The rest features are related to the income, poverty, ethnicity (HISPANIC OR LATINO) and earning of female.

4.2 CORRELATION OF 0.6

When selecting the interesting features with correlation greater than 0.6, most of the selected features are still belong to education data-set. 29 out of 99 are related to household type, family relations and the marital status. 12 features are relevant to poverty. There are also 4 features that belongs to insurance. The rest features are related to the income, poverty, ethnicity and earnings.

4.3 CORRELATION OF 0.5

There are 214 features being selected with the constrain of correlation greater than 0.5. The data-set that are relevant are keep unchanged but more featrues being selected.

The Figure.1 shows that the four most relevant dataset have different types of correlation with the predicted feature (the earning of male over $100000)

4.4 THE TESTING RESULT OBTAINED WITH CORRELATION OF 0.6

The Figure.2 is the earning predicted by the model trained by the 99 features selected with correlation of 0.6.

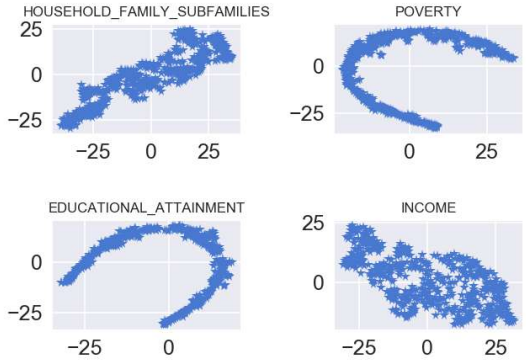

Figure 1: Correlations of selected dataset

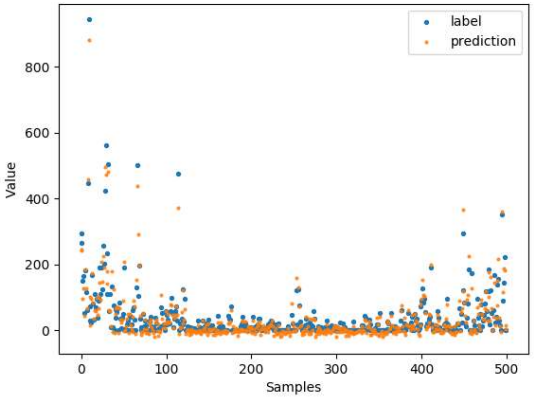

Figure 2: The testing result

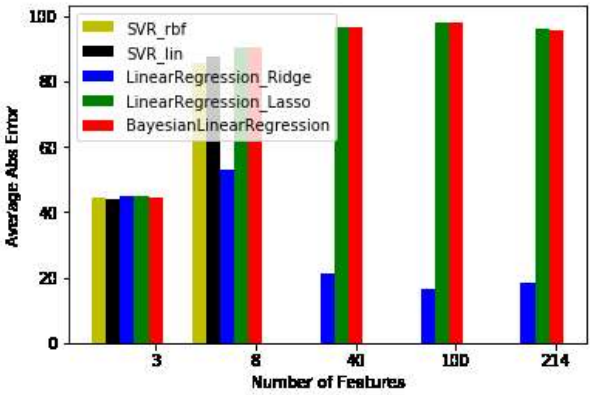

Figure 3: Result

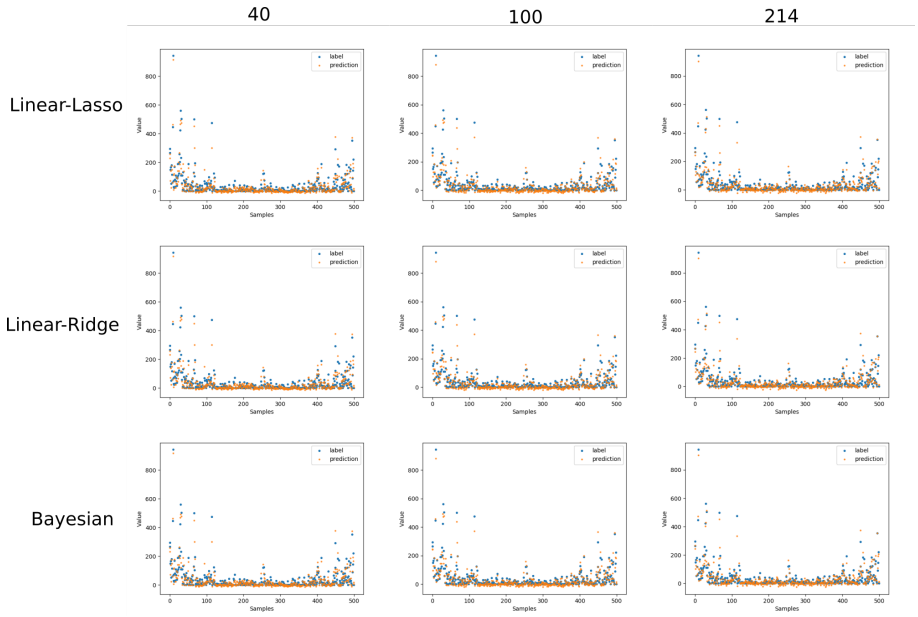

Figure 4: comparison

5 Conclusion & Future Work

In this paper, correlations between different socioeconomic aspects were studied. In particular, we focused on the prediction of high earnings population size. There are five primary factors with relatively strong correlations to the interested feature: household family and subfamilies, education attainment, poverty, income and earnings(females in the family).

It can be evident from not only published literature in a sociologist perspective, but also the calculations of correlations among all pairs of features. The data representing each class were extracted from the whole database. Afterwards those data columns containing ”null” or non-numerical data were eliminated, in order to facilitate the training of the machine learning model. After the pre-processing of data, it was fed into three different types of encapsulated structure in the library SKLEARN, including support vector regression (SVR), linear regression (LR) and Bayesian linear regression (BLR). The training was processed automatically via calling of methods in SKLEARN. Based on the evaluation in validation set, the performance is decent.

Future investigations would be suggested in the field of other features or more robust model structure.

REFERENCES

Samuel Bowles, Herbert Gintis, and Melissa Osborne. Incentive-enhancing preferences: Personality, behavior, and earnings. *American Economic Review*, 91(2):155–158, 2001.

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pp. 155–161, 1997.

David L Featherman and Robert M Hauser. Sexual inequalities and socioeconomic achievement in the us, 1962-1973. *American Sociological Review*, pp. 462–483, 1976.

John Neter, William Wasserman, and Michael H Kutner. Applied linear regression models. 1989.

Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

Gary Solon. Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives*, 16(3):59–66, 2002.