Yong-Chan Shin (SID 25421882)

Prof. Adhikari, Prof. Gonzalez

DATA100

7 May 2020

Contraceptive Exploratory Data Analysis

Abstract

In 1987, the population of Indonesia reached 171.7 million with a rapid growth as fast as that of the United States, and now it reached 4th rank of world population. However, while the United States had maintained the highest GDP, Indonesia's GDP could not even reach 2% of the United States', which means that Indonesians in 1987 could not afford a decent standard of living, suffered from a shortage of commodities, and income per capita could not increase. One way to solve this overpopulation problem is boosting the usage of contraceptives, and the objectives of this exploratory data analysis are finding the features that influence the choice of contraceptives, exploring what other information can be used to predict the number of children of a woman, and predict the contraceptive and number of children based on the other features.

Finding Features Correlated to the Contraceptive method

First of all, let's take a look at the given features; there are "wife_age", "wife_education", "husband_education", … , and "contraceptive", and our target feature, contraceptive, is of no-use, and long-term/short-term categories. Among the rest of the features, "wife_age" and "num_child" are numerical, education levels, occupation, and standard-of-living index are categorical with integer values from 1 through 4, and religion, working status, and media exposure are binary. Since the categorical variables are expressed in numeric values, one-hot-encoding to those variables are required. Then, for instance, instead of "wife_education", new features "wife_education_1" through "wife_education_4" are used, so the new dataset contains two numerical features "wife_age" and "num_child" with 19 binary variables and contraceptive.

Further explorations will use this refined dataset and refer to the Jupyter Notebook figures. Firstly, let's see how many cases fall into each "contraceptive" value; Fig1 is a countplot showing that roughly half of the samples do not use contraceptive, short-term contraceptive method follows the next, and long-term contraceptive method is the least. Next, comparisons between numerical features and contraceptives are described in Fig2 and Fig3 through boxplot. Fig2 gives a first insight that the ages of wives who use long-term contraceptive are concentrated on large values(interquartile range is between 27 and 41 with median 35) while those using short-term contraceptive are concentrated on small values(interquartile range is between 25 and 35 with median 29). However, no-use samples' interquartile range contains the both previous interquartile ranges, so it is expected that if some other feature can classify samples not using contraceptive and those actually using contraceptive, wife age feature classifies long-term and short-term contraceptive methods well. On the other hand, boxplot of Fig3, distributions of the number of children of no-use, is concentrated on smaller values, while the 25th percentiles of the samples using any type of contraceptive are the same with the median of the no-use samples. With very initial intuition, no-use samples are expected to have more children, but, instead, it turns out that no-use samples tend to have a very few children according to the countplot of Fig3.

Now, let's explore categorical variables with low/high levels such as education level and standard living which are of integer values from 1 to 4. These are very interesting features because albeit those are just categorical variables, higher level is expected to have higher employment of contraceptive. Based on Fig4, while education level 1 of wife is dominated by no-use of contraceptive, education level 4 of wife has a very high contraceptive method usage, and especially long-term contraceptive has the largest count which is contrastive with Fig1, the overall contraceptive tendency. Husband education level 4 and standard living level 4 have the same tendency that both the absolute number and portion of both contraceptive methods evidently increased from level 1.

With the features selected, the models to train and predict must be able to do classification. Firstly, decision tree will be used as it can handle both numerical and categorical data. Since there is a high possibility of overfit in decision tree, random forest will be included as well with "n_estimators" parameter as 25 which worked the best among the multiples of 5 from

5 to 30. Logistic regression is designed for classification without any missing values in the dataset, and setting the prediction model as one-versus-rest can handle the three categories of contraceptive independently. “Solver” parameter is set to “liblinear” as the dataset is small enough, “max_iter” parameter is set to 1000 as the number of iterations goes over the default value 100, and regularization parameter “C” is chosen to be 1.0 among the values 0.01, 0.1, 1.0, and 10.0. Stochastic gradient descent classifier will work with several hyperparameters modified; one of the hyperparameters, “alpha”, is the regularization parameter, determined to be 0.1, also chosen among several geometric numbers from 0.00001 to 10.0. Lastly, k-nearest neighbors model is expected to be effective as the refined dataset is mostly of categories, and its core hyperparameter, “n_neighbors” is set to be 25, chosen among multiples of 5 from 5 to 40. All the hyperparameters chosen are based on the cross validation scores, and based on these models, the computations were held with three different cases: Scores without feature selection, scores with the manually selected features, and scores using “SelectKBest” feature selection tool. “SelectKBest” tool is using chi-squared stats as all the feature values are nonnegative, and the core hyperparameter “k” is chosen to be 8 as the highest cross validation score among all the models used goes up until 8 and goes down from 9 among the integers 1 through 10.

Using the “score” function, the training score(t_score) and cross validation score(cv_score) are plotted for every described model and feature selection in Fig5. As illustrated above, a lot of trials with different hyperparameter values are made, but, unfortunately, it was unable to reach 60% cross validation accuracy. The Fig5 says that the random forest and decision tree have very high training scores, but the cross validation scores were around 50% accuracy. In other words, random forest and decision tree are overfitting because of a number of binary variables, while logistic regression, stochastic gradient descent, and k-nearest neighbors are not, but the cross validation scores are still around 50%. Particularly k-nearest neighbors work the best with cross validation score 55%.

Finding Features Correlated to the Number of Children

Now let’s start over with predicting the number of children based on the other information, and skip some repetitive explanations. First of all, let’s see how “num_child” is distributed through Fig6; there are about 100 of “num_child” values of zero, more than 250 of samples have one child, and the rest are decreasing as the number of children increases. Also, it is intuitively predicted that “wife_age” will have some correlation with the “num_child” feature as the younger a woman is, the more likely to have less number of children; the scatterplot of Fig7 shows the expected tendency except that there are still a lot of samples of high “wife_age” but have small number of children. Then how about the other categorical features? Because of a number of redundant graphs, most of the violin graphs are not included, but it was successful to find four categorical variables(“wife_education_1”, “husband_education_1”, “husband_education_2”, and “contraceptive”) which have different “num_child” distributions by its binary values, as plotted in Fig8. Also, “husband_occupation” and “wife_religion” are expected to have strong correlations, but the violin graphs’ distributions look mostly alike, so it turned out to be ineffective.

With these manually selected features, it is possible to do the same scoring procedures with the previous part’s but using a modified function “score2” which do the same job with “score” function but have different hyperparameters of the selected models and two additional models, LASSO and linear regression. Since the “num_child” feature only has 12 different values after dropping values with less than five samples, it is possible to use the previous models used which accepted three different categories in the previous part and now 12. However, since the “num_child” feature is actually a numeric feature, LASSO and linear regression can be introduced this time. Random forest’s “n_estimators” is chosen to be 15 among the multiples of 5 from 5 to 30, and logistic regression’s “C” is set as 20 from 0.1 and multiples of 5 between 1 to 25. Stochastic gradient descent’s “alpha” is chosen to be 0.001 among geometric series of numbers from 0.0001 to 10.0. K-nearest neighbors’ “n_neighbors” are chosen in the same way as the previous part, and LASSO’s “alpha” is chosen among geometric numbers from 0.0001 to 10.0.

Lastly, let’s discuss the performances of the models along with the feature selections through Fig9. As easily predicted, random forest and decision tree are overfitting as the training score is very high while the cross validation score is very low, which is around 20~25%. Logistic regression and k-nearest neighbors did not perform well this time as the target variable is actually not a categorical variable. Thus, the stochastic gradient descent, LASSO and linear regression are expected to perform better, and the result says that LASSO and linear regression performs the best with the cross validation score around 37%. Since “alpha” of LASSO is set to be very small and LASSO with the regularization parameter zero works exactly as linear regression, the result of LASSO is very close to that of linear regression; the difference between the scores do not exceed 3% and mean squared error is also close enough to say so. Although manual feature selection did not perform as well as “SelectKBest” did overall, there are still some models like decision tree, logistic regression, and stochastic gradient descent that performed better.

Conclusion

The biggest challenge for both objectives was that the cross validation scores were not really good; even though a number of trials have been made through almost all models that this course has covered and modified a number of hyperparameters, there was a ceiling that these trials could not break through. However, the number of children with respect to contraceptive was very interesting; there was an intuitive assumption that zero number of children will imply a lower number of non-usage of contraceptive, but, in fact, it was dominated by contraceptive value 0 in Fig3, indicating non-usage of contraceptive. Thus, if additional data is available, the number of sexual relations per specific period will enable more accurate predictions, and any additional numeric variables like age at first birth can also be a confounding factor; in the past, people were more likely to have a child on their earlier age as the life expectancy was much lower than the current's. Another extra helpful data will be how easily can a woman access to a contraceptive; this can be translated to the area information(urban or rural), and, considering the fact that the dataset is from 1987, it is very unlikely to obtain contraceptive online, so the physical distance to pharmacy or medical center will not be negligible. Lastly, despite the analyses based on education level and standard of living, it is very likely to have prejudice about a woman’s contraception based on the other information like education level and standard living. For instance, if the dataset included the area information, it would be helpful for the government to prioritize some regions to build more medical facilities, but, on the other hand, the people living the selected regions are likely to be regarded as less educated people with low standard of living. Also, there may be a chance that there is a religious reason not using any contraceptive method, so governmental intervention to increase the employment of contraceptives may be offensive to the people even if it will definitely help them to have better quality of lives. Still, exploratory data analysis on contraceptive has a huge possibility to guide the government how to control the population through contraceptive, or any other features that have strong correlation with either contraceptive usage or number of children. Not only the additional data mentioned before, but the exact same data in different years will notice whether the government's birth rate policies are going toward the correct way.