Machine Learning Bootcamp - 2020

Team No. 4

Project 1 Final Report

Submitted By-
Pratik Ghatake- Business Analyst
Fatmir Likrama- Developer
Celine Cherian – Data Engineer
Miguel Mendoza – Data Scientist

1. Problem Introduction

In recent years, with the globalized world, it has become even more relevant that governments have started looking at happiness as a matric to measure success. Being happy is a simple yet extremely profound feeling nowadays. Happiness cannot be defined in predefined format. Happiness Index or happiness rank might seem trivial, but it points to the gaping lacunae in the government development policies in each country and one can view the measure as people's perception of how their governments perform. The insights from the analysis of world happiness also helps in pointing out the importance of development qualitatively rather than quantitatively.

We have focussed on the dataset from the World Happiness Report Landmark Survey. Furthermore, we focus on predicting the happiness score of the country based on the independent features. Also, we have dealt with bias and variance in the dataset with regularization thereby optimizing the performance of predictive model. The other aims of the analysis include identifying the key variables for happiness, analysing the countries which are consistent, maintaining good happiness score, etc.

2. Dataset

The dataset used for this project is published by the United Nations. The United Nations publishes the World Happiness Report every year. As far as the boundaries of the analysis are concerned, we have circumscribed the report to years starting from 2015 to 2019. The rankings of the happiness report are based on a Cantril ladder survey**.** Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The report correlates the results with various life factors. There are different fields involved in the dataset including economics, psychology, national statistical figures etc. which are measured on different scales and are used to effectively assess the happiness score of the country. There are a total 782 observations of different countries for the span of 5 years in total with 9 different variables.

3. Features and Processing

Without domain knowledge, we could not decide which attributes are potentially informative. There are a total 5 datasets, one for each year from 2015 to 2019. The datasets were merged into a

single dataset and concerned pre-processing is carried on the dataset in order to process smooth merging. Data pre-processing in this project involves achieving data uniformity by defining variables, rearranging the columns, adding new columns etc. Columns following the happiness score take into consideration six factors which make life evaluations higher in each country.

The dataset contains the 9 variables including the target variable and 782 instances. In the data preparation and processing phase, we looked upon different features and their linearity with the Target data. We also checked if the data possess any multicollinearity issue. The distribution of the features is checked using matplotlib for pinpointing the outliers or skewness in the data. There was only 1 missing value in the dataset which processed further with mean replacement. Since, we were not having any outliers we chose mean replacement. At the end, the features were normalized on the same scale to improve validation accuracy.

4. Models and Techniques

- **Ridge Regression**: Ridge multilinear regression was the first model used to estimate our coefficients. This technique uses a typical multilinear regression equation and adds a penalty parameter, equal to the sum of the squares of the coefficients multiplied by an alpha constant. This parameter is added to reduce the chance of over-fitting. This model was implemented in Python through *sklearn's* Ridge and *RidgeCV* module.

- **Lasso Regression**: Like Ridge regression, this model is a modification of the typical linear regression equation, adding a penalty parameter. However, this parameter uses the absolute value of the coefficient, or L1 norm, allowing coefficients to go to zero if they do not influence the predicted score. This model is implemented like the ridge regression, using *sklearn's* lasso and *lassocv* module.

- **Elastic Net Regression:** To achieve benefits of both ridge and lasso algorithms, elastic net regression was performed with multiple values of alpha and l1 ration ranging from 0 to 1. As per the requirements of this algorithm, features were normalized before processing them for this algorithm. Using *GridSearchCV* techniques hyperparameter tuning is performed from which the selected best value of *alpha* and *l1 ratio* is 0.001 and 0.8 respectively.

5. Results

- **Ridge Regression**:

  From the ridge regression, we observed the coefficient estimates showed inverse relationship with the value of alpha. The mean squared error of the prediction is 0.29. Since the cross-validation techniques were applied on the algorithm, it resulted into low mean squared error and lessened the chances of overfitting.

  The techniques used with multiple algorithms allowed us to get a wide understanding of how we would e able to evaluate the and deploy the findings. The adjoined graph shows the scatterplot of prediction error. Since none of the coefficients are reduced to 0, we can infer that the algorithms did not perform variable selection.

- **Lasso Regression**: Lasso algorithm was deployed with *max_iterations = 10,000* and 10-fold cross validation to choose the best value of alpha. The mean squared error was little worse than the test MSE of ridge regression with alpha chosen by cross validation. The dataset has group of variables with very high pairwise correlations because of which lasso arbitrarily selected only some of the variable from the group. The scatterplot shows the actual vs Predicted values.
- **Elastic Net Regression:** The coefficients of the independent features are almost the same. *Freedom to make life choices feature SQ* was completely removed by the algorithm since it showed high correlation with other variables. The *best_score_* selected was 0.35 with MSE as 0.30.