Sensor Values Prediction: The Exploration of Regression Models from Educational Sites and Research Literature

Raymond Liu
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada

Ravishankar Subramani Iyer
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada

ABSTRACT:

Regression refers to the technique of modelling data which is continuous (measurable) in nature. To find the best regression model to predict sensor value, we explored multiple regression models from StackOverflow and research literature to make quantitative predictions of the failure scores of 103 sensors. From the 21 different regression models, GBDT performed the best in terms of Root Mean Squared Error (RMSE) (17.94 ± 2.62) and R2 scores (0.17 ± 0.22). Subsequently, we optimised the GBDT model to predict the failure scores of the 103 sensor values of the subsequent year. This project not only considers almost every possibility of regression prediction use cases but also ensures the coverage of the possible appropriate regression model to perform the best result.

KEYWORDS

Regression, Statistical evaluation, Modelling, Machine learning, Prediction

1. Introduction and Background Work

Regression refers to the technique of modelling continuous data. In the real world, we come across these data types to predict future or subsequent quantities such as weather conditions, stock market predictions, and house prices. Given the rapid advancements in machine learning approaches to solve real-life problems such as prediction, on the other hand, research literature majorly considers the traditional algorithms to train, optimise and predict their data. Optimistically, several researchers have attempted to implement advanced or state or art algorithms. For example, Wang & Cao [1] used Linear Regression, Decision Trees, and Random Forest to predict office building rental costs, while Shanti et al [2] used Artificial Neural Networks (ANN). In addition to traditional algorithms such as Support Vector Regression (SVR), we also see a variation in the model selection approach. Roziqin et al [3] used the Monte Carlo version of Linear Regression and Polynomial Regression to predict dengue fever cases, whereas Yang and Cao [4] utilised an ensemble learning approach to predict house prices.

Given the previous study, we explored the implementation of state-of-the-art algorithms from educational websites and traditional algorithms from a literature review to predict the failure scores of 103 sensor data of the subsequent year.

2. Methodology

The training examples contained 103 sensor values for 208 consecutive weeks, and the output value contained 208 failure scores from the sensors corresponding to each of the 208 weeks. Since the

output values are continuous, we consider training the regression algorithms to predict the sensor readings failure score of the subsequent year.

The approach for selecting regression models was based on a summary of the literature review (in the domain of predictive modeling) and educational websites (such as StackOverflow). We trained all the selected models using their respective default parameters by employing 10-fold cross-validation. The evaluation was performed by commonly used regression metrics such as the *mean-squared error loss function* between the given output value and the predicted output value, the *Root Mean Squared Error* (RMSE), and the *R2 score*. Subsequently, we considered 5 models for optimization. The performance results are indicated in the supplement file.

Next, we optimized each of the 5 models with suitable parameter values using 10-fold cross-validation and obtained mean-squared error loss function values for each of the values. Further, we derived their learning curves of *Mean Squared Error* (*MSE)* and *R2 score* evaluation metrics. The best-performing model out of the five models was used to predict the sensor values of 103 sensors of the subsequent year.

3. Results

According to the basic statistics of the training dataset, the median outcome is 115.91, and its lower and higher quartiles are 96.58 and 136.03, respectively.

Overall, we considered 8 regression models from the literature review and 13 models from StackOverflow (See the supplement file). We evaluated each of the model's performances using *RMSE* and *R2 scores*. According to the result, five models, including Random forest (RF), XGBoost, Bayesian Ridge, ARD Regression, and GBDT performed the best.

After optimising the parameters of each model, we found that the GBDT model performed the best among the five, both in terms of RMSE (17.94 ± 2.62) and R2 scores (0.17 ± 0.22) as shown in Table 1. Therefore, we used the *GBDT model* to predict the sensor values of 103 sensors of the subsequent year.

Table 1. Cross Validation Results of the final five Models

Model	RMSE ± 1.96 * std	r2 score ± 1.96 * std
RF	18.59 ± 2.46	0.12 ± 0.23
XGBoost	18.28 ± 2.49	0.15 ± 0.18
Bayesian Ridge	20.73 ± 2.49	-0.13 ± 0.29
ARD Regression	22.83 ± 2.87	-0.42 ± 0.39
GBDT	**17.94 ± 2.62**	**0.17 ± 0.22**

Abbreviation: ARD: ARD: Automatic Relevance Determination; GBDT: Gradient-boosting decision trees; RF: Random Forest; RMSE: Root Mean Square Error; XGBoost: Extra gradient boosting.

4. Discussion and Conclusion

The initial selection of models based on their default parameters is effective since we could filter the inappropriate models for the dataset. It was implemented by monitoring the distribution of prediction results and cross-validation results such as RMSE, and R2 scores. Some models, such as RANSAC Regressor, SGDRegressor, Linear Regression, and Lars could be excluded directly due to the unreasonably large prediction values (here, Linear Regression) and outlying cross-validation results (here, RANSAC Regressor, SGDRegressor, and Lars). Despite the prediction and

cross-validation accuracy of other models closer to each other, we chose five models that performed the best by default. This selection not only improved the overall efficiency but also enabled us to focus on the model that potentially fit the data best.

Despite the optimization, the mean R2 score in cross-validation results tended to be very low and did not elevate significantly. It is probably due to the distributional difference between the prediction value (usually greater than 120) and its ground truth (median = 115.91). However, the learning curve (see Colab notebook) showed that the R2 score could be as high as around 0.6 for the selected model when the training size increased to nearly 80% of the entire dataset, indicating that the R2 score is very prone to the distribution difference between the predicted result and the ground truth. Despite this, for the model whose distribution of prediction value is closer to its ground truth, such as TheilSenRegressor, its RMSE and R2 score were poorer than many other models (see supplement file), leading to its elimination. Such a scenario indicates that the same distribution between ground truth and the result does not guarantee a more precise prediction due to the excessive individual variance between the prediction and the result. Additionally, the model we found in the literature review [1-4] tended to be better than those from StackOverflow, indicating that the reliability of finding relevant models depends on the specific problem and, in addition, the amount and quality of data we have for that problem.

There are several limitations of this project. First, the sample size is limited, leading to the limited performance of the model and the non-convergent learning curve. Second, the distribution of the prediction result is generally greater than the corresponding ground truth, leading to the limited accuracy of the model. As a future scope, there could be several regression models worth exploring. Adding to the expansion of the stochastic nature of the dataset and the measures of improving its representability, we can achieve the degree of model usage and optimization with the capability of better fitting the dataset to the real-time scenario.

Conclusively, this project explored multiple regression models from educational sites (StackOverflow) and research literature to make quantitative predictions. By exploring a wide range of different models from different resources, this project not only considers almost every possibility of regression prediction use cases but also ensures the coverage of the possible appropriate regression model to perform the best result.

REFERENCES

[1] Zhihong Wang and Buyang Cao, “Prediction of Office Building Rental upon Spatiotemporal Data,” 2019 2nd International Conference on Data Science and Information Technology (DSIT), Seoul, Republic of Korea, 2019, pp. 168–174, doi: 10.1145/3352411.3352438.

[2] Naser Shanti, Akram Assi, Hamza Shakhshir and Adnan Salman, “Machine Learning-Powered Mobile App for Predicting Used Car Prices,” 2022 3rd International Conference on Big-data Service and Intelligent Computation (BDSIC), Xiamen, China, 2022, pp. 52–60, doi: 10.1145/3502300.3502307.

[3] M. C. Roziqin, A. Basuki and T. Harsono, "A comparison of Montecarlo linear and dynamic polynomial regression in predicting dengue fever case," 2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC), Manado, Indonesia, 2016, pp. 213-218, doi: 10.1109/KCIC.2016.7883649.

[4] Y. Bowen and C. Buyang, “Research on Ensemble Learning-based Housing Price Prediction Model,” 2018 Big Geospatial Data and Data Science, 2018, pp. 1-8, doi: 10.23977/bgdds.2018.11001.