Prediction of Lung Cancer Based on Catboost

Leqi Ma[1, *, #], Yimeng Lu[2, #], Yiduo Liu[3, #]

[1]School of Mathematics, Norwest University, Xi'an, China, 710127

[2]Department of Industrial Engineering, Tsinghua University, Beijing, China, 100084

[3]School of Environment, Beijing University of Agriculture, Beijing, China, 102206

* Corresponding Author Email: mlqi0806@163.com

#These authors contributed equally.

Abstract. This research aims to provide guidance for achieving precision medicine by accurately predicting the incidence of lung cancer. This paper used random forest screening to identify several variables that make significant influences on the lung cancer, and used Smote oversampling to address the issue of data imbalance. Finally, this paper used Catboost to construct a model to handle categorical features. Through analyzing experimental data, can found that among the selected variables, age has the greatest importance. Afterwards, validated the model and the training score was 0.9032, indicating good results. At the same time, establish a confusion matrix to prove once again that the model has good predictive performance. Finally, based on cross validation, both the optimal validation accuracy score and the accurate validation accuracy score were above 0.9, indicating that the model performance is excellent. The innovation point of this paper is to realize precision medicine through high accuracy prediction of lung cancer.

Keywords: Random Forest, Smote Oversampling, Catboost.

1. Introduction

1.1. Research background

Current status of lung cancer according to the 2019 World Health Organization, cancer is the first or second leading cause of death in more than 57% of countries, and the third or fourth leading cause of death in more than 12% of countries. Cancer is not only one of the main reason to causes the death ofpopulation, but the treatment of cancer will be a huge medical burden, posing a great public health challenge [1].

1.2. The purpose of this research

This research focuses on understanding the current situation of lung cancer and the relationship between multiple variables and the incidence of lung cancer, and to explore the differences in the influence of the same variable on the incidence of cancer among different groups. Through horizontal comparison of various behaviors that may lead to cancer and the possible symptoms of lung cancer in a specific population, the variables with high correlation with lung cancer patients were found out, and the possible causes of cancer were explored. Therefore, this project has guiding significance on how to correctly prevent lung cancer, put an end to bad life behavior and screen early symptoms of lung cancer [2-4].

2. Data sources and exploratory analysis

2.1. Data sources

As shown in Table.1, there are 16 variables in the data, of which age is a continuous variable and the remaining variables are categorical variables [5].

Table 1. Data Variables Summary

Serial number	Variable Name	Description of the variable
1	GENDER	Gender
2	AGE	Age of patients
3	SMOKING	Smoking yes =2 no =1
4	YELLOW_FINGERS	Yellow fingers yes =2 no =1
5	ANXIETY	Anxiety is =2 no =1
6	PEER_PRESSURE	Having peer pressure is =2 no =1
7	CHRONIC DISEASE	Chronic disease yes =2 no =1
8	FATIGUE	Fatigue yes =2 no =1
9	ALLERGY	Allergy yes =2 no =1
10	WHEEZING	Wheezing yes =2 no =1
11	ALCOHOL CONSUMING	Drinking yes =2 no =1
12	COUGHING	Cough yes =2 no =1
13	SHORTNESS OF BREATH	Shortness of breath yes =2 no =1
14	SWALLOWING DIFFICULTY	Dysphagia is =2 no =1
15	CHEST PAIN	Chest pain yes =2 no =1
16	LUNG_CANCER	Lung cancer yes =2 no =1

2.2. Exploratory analysis

As shown in Figure 1, in the statistical data, 52% males and 48% females, the ratio is close to 1:1, and there is no excessive deviation.

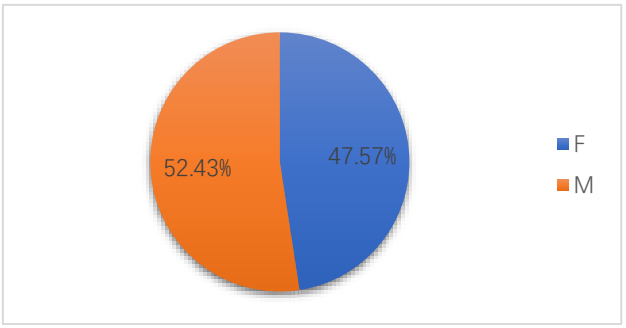

Figure 1. Pie chart of gender distribution

In terms of age proportion, there are 97 people in the range of 60 to 67 years old and 92 people in the range of 54 to 60 years old. As shown in Figure 2, the boxplot of the variable AGE shows that the average age of men and women is almost around 65, with few abnormal values and extreme values concentrated in women, with the youngest being 21 years old and the oldest being 87 years old.

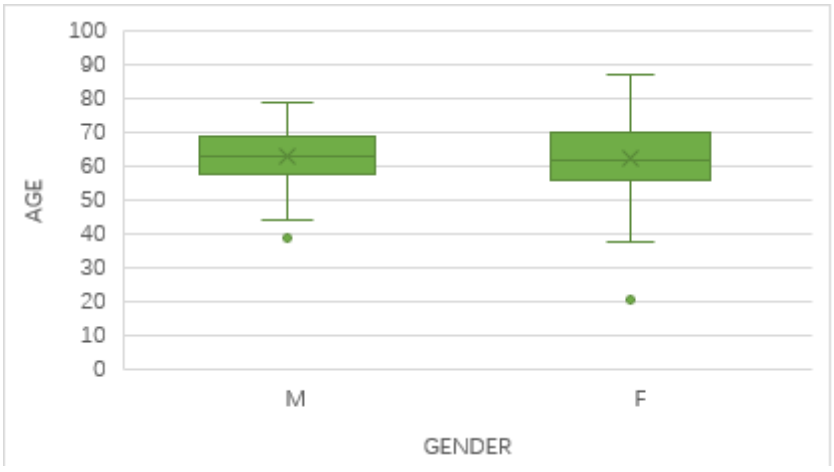

Figure 2. Box plot of age distribution for men and women

As shown in Figure 3, in variable 16, 87.4% of the subjects were diagnosed with cancer, and 12.6% of the subjects did not suffer from lung cancer although they had various symptoms to varying degrees.

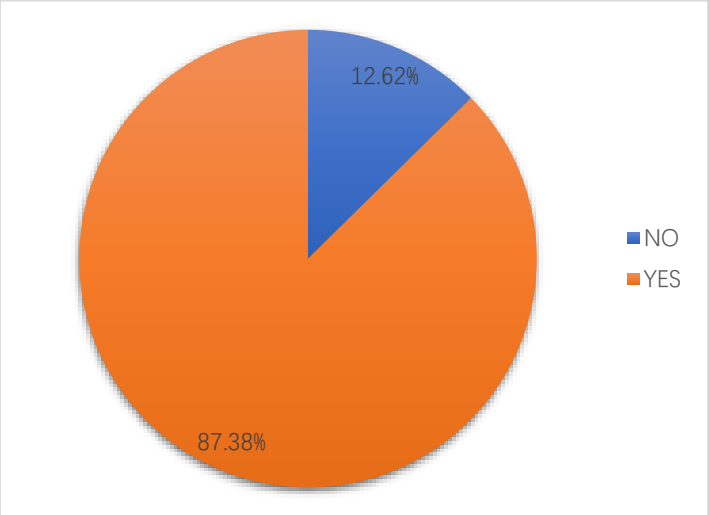

Figure 3. Pie chart of whether or not the patient has lung cancer

3. Methods

3.1. Principle analysis of pretreatment (normalization)

As we want to eliminate and the dimensional influence between those different indicators, data standardization is implemented to solve the comparability between those data indicators. The most typical way is the normalization of data. In this study, Min-Max Normalization method was used.

3.1.1 Principle of Min-Max Normalization algorithm

This normalization method is to transform the original data into the range of [0,1] by linear function, and the calculation result is the normalized data. The calculation formula is as follows:

$$x' = \frac{x - \min(x)}{max(x) - \min(x)} \quad (1)$$

Where x' denotes the transformed data, x denotes the original data, $min(x)$ is the minimum value in the original data, $max(x)$ is the maximum value in the original data.

3.2. Principle analysis of random forest feature screening

Random forest is one of the ensemble learning bagging class methods and one of the earliest ensemble learning algorithms. It acts as a bagging class method, it can show better performance than decision trees alone on almost the vast majority of datasets, while it can also be used as a feature selection method in its own right. Its principle is described specifically in [9] and [10].

3.2.1 Feature selection of random forest

Based on the method of feature quality measurement, the steps of feature selection in random forest are as follows:

(1) The first step is to calculate the importance of each feature, and then sort it in descending order.

(2) The proportion to be removed was determined, and the corresponding proportion of features was removed according to the importance of the feature to obtain a new set of features.

(3) Repeating the above process with the new feature set until there were remaining features (a was the value set in advance).

(4) According to each feature set obtained in the above process and the corresponding out of-pocket error rate of the feature set, then we select the feature set with the lowest out-of-pocket error rate.

3.3. Smote oversampling principle

In various models, sample imbalance is a common issue, that is, the proportion of positive and negative samples is seriously out of balance. In such cases, traditional machine learning algorithms may tend to bias the dominant category, resulting in degraded model performance. To solve this problem, a commonly used approach is oversampling, where Smote oversampling is a widely used technique. Its principle is shown specifically in [11] and [12].

3.3.1 Smote oversampling algorithm steps

Smote oversampling method is based on the feature space of samples, and generates synthetic samples by interpolating the minority samples. The main steps of SMOTE are as follows:

(1) For each sample x in the minority class, the Euclidean distance is used as the standard to calculate its distance to all samples in the sample set of the minority class, and its k-nearest neighbors are obtained.

(2) For each minority class sample x , a number of samples are randomly selected from its k nearest neighbors, assuming that the selected nearest neighbors are x_n.

(3) For each nearest neighbor x_n, constructing a new sample with the original sample by the formula:

$$x_{new} = x + rand(0,1) \times (\tilde{x} - x) \tag{2}$$

Where x_{new} is the constructed new sample, x is the original minority sample, $rand(0,1)$ is the random number with values in the range of 0 to 1, $\tilde{x}$ is the randomly selected nearest neighbor x_n.

4. Catboost principle analysis

Catboost is an open source library that combines boosting, with a new algorithm for handling categorical features. In the following, we will analyze and introduce the algorithm principle. Its principle is shown specifically in [7] and [8].

4.1. Category features

A category-type feature is a feature that has a discrete set of values, called categories, that are not comparable to each other. In the case of high cardinality features, such as user ids, this approach leads to a large number of infeasible new features. To solve this problem, categories can be grouped into a limited number of clusters and then one-hot encoding can be applied, while a popular approach is to group categories according to the target statistic (TS), which estimates the desired target value in each category.

4.2. Ordered Target Statistics

A more efficient strategy is used in Catboost. It relies on a ranking criterion and is inspired by online learning algorithms that acquire training examples in chronological order. For a given sample, the value of TS depends on the observation history. In order to apply this idea to offline data, the algorithm randomly sorts the data. For each sorted sample, the TS value of the sample category is calculated using the previous data of the sample, and multiple random sequences are used to reduce the variance of TS value.

4.3. Feature combination

Another important implementation of Catboost is to use combinations of different categories of features as new features to obtain high-order dependencies. Catboost adopts a greedy strategy to achieve the combinations. For the first segmentation of the tree, no combination is applied. For the next segmentation, it combines all combinations and categorical features of the current tree with all categorical features in the dataset, and dynamically transforms the categorical features obtained by the new combination into numerical features.

4.4. Predict offset/gradient bias

To solve the problem of gradient bias, Catboost makes some improvements to classical gradient algorithms. In Catboost, in the first stage, it adopts the unbiased estimation of the gradient step size, and then uses the traditional scheme named GBDT to solve the gradient bias.

4.5. Advantages of Catboost

Catboost offers many advantages, such as superior performance (it rivals any advanced machine learning algorithm in terms of performance), robustness (it reduces the need for much hyperparameter tuning and reduces the chance of overfitting).

5. Results

5.1. Data cleaning and preprocessing

When processing the existing data, a variety of factors need to be taken into account, so the existing data should be pre-processed. Data preprocessing refers to the reasonable clarity and preprocessing of the data before modeling or mining the existing data. It mainly includes the processing of missing data, coding, normalization, deleting invalid features, etc., so that the data meets the conditions, and then modeling or data mining is performed to improve the efficiency of the model. There are no missing values in the data used in this paper, so the data.

5.1.1 Feature coding and normalization

Since most classifiers cannot perform direct processing for typed variables, coded processing is performed for typed variables. In this paper, the "GENDER" categorical variable is coded as 1 for "MALE" and 0 for "FEMALE"; For the remaining 13 categorical variables, the code for "yes" was 1, another code for "no" was 0. The only continuous variable, age, is normalized to map data to a range so that measures of different units or magnitudes can be compared and weighted and calculated. All things considered,Min-Max Normalization is selected to map the data between [0,1].

5.1.2 Feature screening

Feature selection can reduce the number of unrelated variables, thereby reducing the problem of model overfitting. On the basis of data preprocessing, this paper selects the random forest algorithm for feature extraction of 16 variables (including the target variable). Random forest is an algorithm containing multiple decision trees, which can train and predict samples. It is mainly used in classification and regression problems. Through the Python programming language, the feature importance ranking can be obtained as shown in Figure 4.

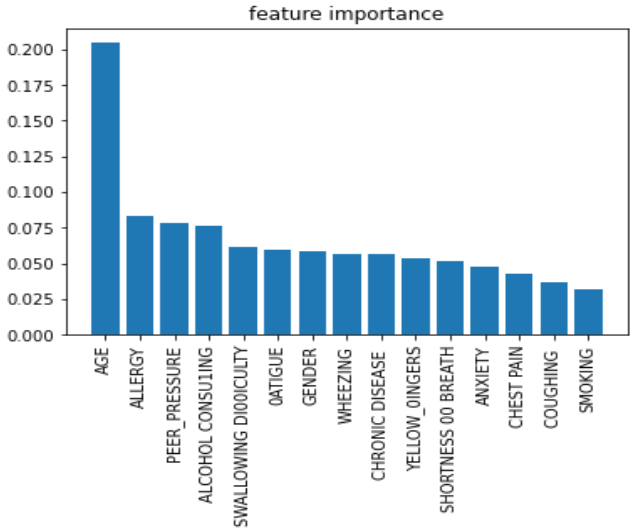

Figure 4. Feature importance ranking

With the threshold set at 0.05, a total of 11 variables were selected, among which "AGE" was the most important feature.

5.2. Ubalanced sample treatment

Sample imbalance will affect the learning process of the classifier. During the exploratory analysis of the data, it was found that the sample size o0000f the target variable was significantly different from that of the sample size of the non-cancer sample, which needed to be processed for sample imbalance. In this paper, SMOTE oversampling method is used to process the data, and new data are synthesized on the basis of the original sample data, which is a sampling method to increase the minority data.

Results before and after treatment were shown in Table.2:

Table 2. SMOTE oversampling results

Label	Before treatment	After processing
0	24	192
1	192	192
Total amount	216	384

5.3. Model training and result analysis

On the basis of feature screening and sample non-processing of the data, Catboost algorithm is used to model the data. The results of model training are shown in Table.3:

Table 3. Results of model

Best Test	Best Iteration
0.9032	398

It can be seen that the best test result of the model is 0.9032, which is higher than 0.9, indicating that the model established has a good fitting effect. In order to better understand the connotation of this model on the validation set, a confusion matrix was established to measure the prediction effect of the model. Confusion matrix is a common index used to evaluate the classification model. The actual category is compared with the model predicted category to evaluate the effect of the model prediction. Four basic indicators (first level indicators) used in the confusion matrix:

- TP (True Positive): The true class of the sample is positive, and the result predicted by the model is also positive.
- FP (False Positive, false positive) : The sample's true class is negative, but the model predicts it to be positive.
- TN (True Negative) : The true class of the sample is negative, and the model predicts it to be negative.
- FN (False Negative, false negative) : The true class of the sample is positive, but the model predicts it to be negative.

When these four metrics are presented together in a table, we get a matrix like Table.4, which we call the Confusion Matrix:

Visualize the confusion matrix, as shown in Figure 5

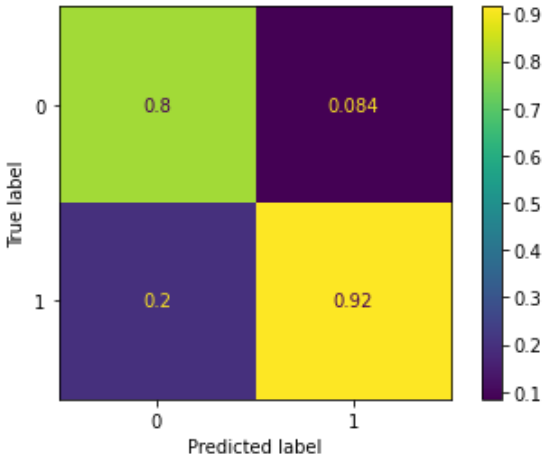

Figure 5. Confusion Matrix

Through the confusion matrix obtained by the model, it can be seen that the corresponding values of the observed values in the second and fourth quadrants are 0.8and 0.92. Conversely, the observed values at the corresponding positions in the first and third quadrants are 0.2 and 0.084. The prediction accuracy of positive and negative instances of the model was 92% and 80%, respectively. It can be seen that the prediction effect of the model is relatively good. In order to verify whether the model has the best performance, cross validation is carried out. K=3 was selected for three times of training and testing, and the results were shown in the following table. The best test results were all above 0.9, indicating that the model had good prediction results:

On the basis of cross-validation, we can obtain that loss function value of the average of 3 folds for each boosting step, which provides us with a more accurate estimate of model performance. The final score we obtain is the average of 3 times results, and the results are shown in Table.4:

Table 4. Final Score

Best validation accuracy score	Precise validation accuracy score
0.91±0.01 on step 139	0.9094

Accuracy is the index used to evaluate the classification model, that is, the probability of predicting the correct rate in all samples. The average classification accuracy can be used as the performance index of the classifier or model. Table 6 shows that the average best validation accuracy score and the exact validation accuracy score are above 0.9. It shows that our model performances well.

6. Summary of the article

Through the analysis of the research status of lung cancer and the research value of lung cancer prediction,and the obtained lung cancer data were analyzed and the Catboost algorithm was used to establish a prediction model. The lung cancer screening can reduce the mortality of lung cancer patients significantlly, and individuals with high risk of cancer can greatly benefit from the screening and reduce the harm. It can also help doctors to assess the risk of cancer in individual patients, and then make better decisions, and has guiding significance for the screening and prevention of lung cancer patients.

References

[1] Liu Z C, Li Z X, Zhang Y, et al. Interpretation of 2020 Global cancer statistics report [J]. Journal of Integrative Oncology Therapy,2021,7(02):1-14.

[2] FERLAY J, ERVIK M, LAM F, et al. Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer[Z/OL]. [2021-02-20]. https://gco.iarc.fr/today.

[3] Yin J, Wang M Y, YOU W C, et al. Interpretation of cancer statistics in the United States and comparison of cancer incidence in China and the United States in 2022 [J]. Journal of Integrative Oncology Therapy, 2021,8(02):54-63.

[4] Li Yuting, Wan Shaoping, Yang Zhonghua. Research progress on influencing factors of lung cancer screening behavior [J]. China Public Health,2023,39(04):521-525. (in Chinese)

[5] Hong, Z.Q. and Yang, J.Y. "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane", Pattern Recognition, Vol. 24, No. 4, pp. 317-324, 1991

[6] Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 1-45

[7] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.

[8] Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset[J]. International Journal of Computer and Information Engineering, 2019, 13(1): 6-10.

[9] Speiser J L, Miller M E, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert systems with applications, 2019, 134: 93-101.

[10] Paul A, Mukherjee D P, Das P, et al. Improved random forest for classification[J]. IEEE Transactions on Image Processing, 2018, 27(8): 4012-4024.

[11] Maldonado S, López J, Vairetti C. An alternative SMOTE oversampling strategy for high-dimensional datasets[J]. Applied Soft Computing, 2019, 76: 380-389.

[12] Elreedy D, Atiya A F. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance[J]. Information Sciences, 2019, 505: 32-64.

[13] Seibert R M, Ramsey C R, Hines J W, et al. A model for predicting lung cancer response to therapy[J]. International Journal of Radiation Oncology* Biology* Physics, 2007, 67(2): 601-609.

[14] Banerjee N, Das S. Prediction lung cancer–in machine learning perspective[C]//2020 International Conference on Computer Science, Engineering and Applications (ICCSEA). IEEE, 2020: 1-5.

[15] Doppalapudi S, Qiu R G, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches[J]. International Journal of Medical Informatics, 2021, 148: 104371.