

Competition Title: Social insurance anti-fraud analysis based on artificial intelligence

Industrial Applications ☐Intelligent Manufacturing ☐Intelligent Driving ☐Intelligent Life ☐Smart Medicine ☒Smart City

[Overall background]

The large scale of social insurance data in China has covered hundreds of millions of people, including the old-age, unemployment, medical, industrial injury, birth and so on, with typical big data features, and will become an important foundation for the development and innovation of intelligent medical treatment.

In recent years, with the promotion of the basic realization of medical insurance and the progress of instant settlement, the situation of medical insurance service has also appeared, and the phenomenon of fraud and fraud has increased. The human society department has achieved a definite effect by using data comparison and screening of diagnosis and treatment rules, but the application of data is still more traditional.

This item aims to promote the application of big data, artificial intelligence and other technologies in the intelligent monitoring field of medical insurance for basic medical insurance, to improve the pertinence and effectiveness of medical insurance intelligence monitoring, so as to improve the ability of social security services and levels in an all-round way. To form a batch of application models and construction plans that can be used for reference, replicable and popularized.

[Business background]

At present, in view of the increasing number of fraud and fraud in medical insurance, the human and social departments have strengthened the monitoring of all kinds of medical services such as out-patient, hospitalized, medicine and other medical services by means of data comparison and diagnosis and treatment rules, but it is still unable to fully and effectively curb the occurrence of such behavior. This item aims to pass big data and artificial intelligence. And other technologies to change the intelligent medical monitoring system, improve the technical ability and level of accurate identification of fraud and fraud, improve the efficiency of medical insurance service automation, and thus build a more fully, more comprehensive, more balanced social security system.

Project description

[Problem description]

The participant team should complete the development and design of the data algorithm model of the team, realize the accurate identification of the fraud and irregularities of all kinds of medical insurance funds, in order to further enrich the medical insurance rules and medical rules of the current medical insurance intelligence monitoring, and improve the pertinence and effectiveness of the medical insurance intelligence monitoring. Examples of irregularities are as follows:

- (1) In order to obtain unsuitable interests, some members collect social security cards from medical insurance personnel with various ways, and carry out false diagnosis and treatment through the social security card to the hospital and cheat medical insurance funds.
- (2) In the diagnosis and treatment of special diseases in out-patient department, some people cheat medical insurance funds through making up medical records and diagnosis and treatment process.

In this competition, the above two kinds of offenders are collectively referred to as suspected fraud. The participants need to obtain the model based on the given training set data, and then apply the model to determine whether the personnel in the testing center are suspected of fraud.

[User expectations]

1. Through the above data algorithm model, the future application of each basic medical insurance service can automatically predict the probability of fraud. On this basis, a group of high fraud probabilities applications are automatically rejected, the extremely low fraud probability application is automatically passed, and the other remaining applications are introduced to the manual examination.

2. Through the above data algorithm model, we analyze the mode of user fraud, so as to improve and enhance the current social security monitoring rules screening system.

[Expected economic effect]

According to statistics from domestic news websites, billions of Yuan can be recovered every year.

[Competition arrangement]

(1) After the registration is successful, the team downloads data samples, local debugging algorithm, and submits the prediction results of the model algorithm. If the team submitted the result many times in one day, the new version will cover the old version. Registration and training test data download website will be announced later.

(2) From x month, x day, the platform carries out a daily evaluation and ranking, and the start time is AM 10:00 on the same day. The ranking will be updated from high to low according to the evaluation index, and the list will be updated every day.

(3) This competition list uses the A/B list, uses 50% of the test set as the A test set, and the other 50% is the B list test set, and the final online results and rankings are scores and ranking on the list of B (the B list is only published on the last day of the resume).

(4) The deadline for competition is x month, x day AM 10: 00.

Task requirements

[Technical path]

Comprehensive use of machine learning, deep learning, integrated learning, statistics and other fields.

[Technical indicators]

Accuracy (precision), recall rate (recall) and F1 value for this item are used as evaluation indexes. The specific calculation formula is as follows:

$$\begin{aligned}\text{Precision} &= \frac{|\cap (\text{PredictionSet}, \text{ReferenceSet})|}{|\text{PredictionSet}|} \\ \text{Recall} &= \frac{|\cap (\text{PredictionSet}, \text{ReferenceSet})|}{|\text{ReferenceSet}|} \\ \text{F1} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\end{aligned}$$

among them, PredictionSet is a collection of suspected counterfeiters predicted by the algorithm, and the ReferenceSet is the true answer to the suspected counterfeiters. The F1 value is used as the final evaluation criterion.

[Submission standard]

The team will predict the test set based on the training data model algorithm and submit the prediction results.

1. Predicting the number of suspected fraudulent personnel using 1 to indicate, and those who are not suspected of fraud using 0 to indicate; the wrong expression will affect the evaluation results.
2. Predicting in the second column according to the staff order in the first column in the format.
3. Being careful not to include column names in the submitted answers (consistent with submitted format CSV).
4. Needing to submit the prediction results to all the personnel in the test center. If the personnel submission is less than or larger than the number of the test set, the results will not be provided.
5. For the submission, the first column ID and the second column forecast tag should use the comma to separate and submit the CSV (Comma-Separated Values) format.

[Task list]

Test set prediction results CSV file.

Reference information

[Reference tool]

Tensorflow, pytorch, xgboost, lightgbm, sklearn

[Reference data]

Machine learning methods, statistical learning methodology, etc.

[Data interface]

The data samples of this contest are the data of medical insurance settlement desensitization of medical insurance in some areas for the past year, including the records of medical expenses and the details of cost details, which have been desensitized and do not contain privacy information.