2021-4-26

Prediction of Outcomes of COVID-19 Cases

CMPT 459 E100 - Data Mining

Course Project Report

Professor: Dr. Martin Ester

Group Members: Yuqing Wu, Zixi Bai, Zeyong Jin

GROUP NAME: FULL MARK 459

Contents

Project Title

The title of this course project is “Prediction of Outcomes of COVID-19 Cases”.

Problem Statement

The project is aimed to predict the outcome of a case. First, the group does the data preprocessing. Then, the group builds 3 models and uses models to predict the outcomes on cases_test.csv dataset and finds out which model is the best one. The performance of these models will be evaluated based on the predictions on the test set.

Dataset Description and EDA

The data set of this task mainly includes' age ',' sex ',' province ',' country ',' latitude ',' longitude 'and' date '_ confirmation', 'source', 'outcome', 'area_ last_ update','area_ confirmed_ cases', 'area_ deaths_ cases', 'area_ recovered_ case', 'area_ active_ cases', 'location_ info', 'area_ incidence_ rate','area_ case_ fatality_ The ratio 'field describes some basic information. Through this basic information to predict the basic situation of patients.

The 4 major visualizations of dataset are shown as follows.

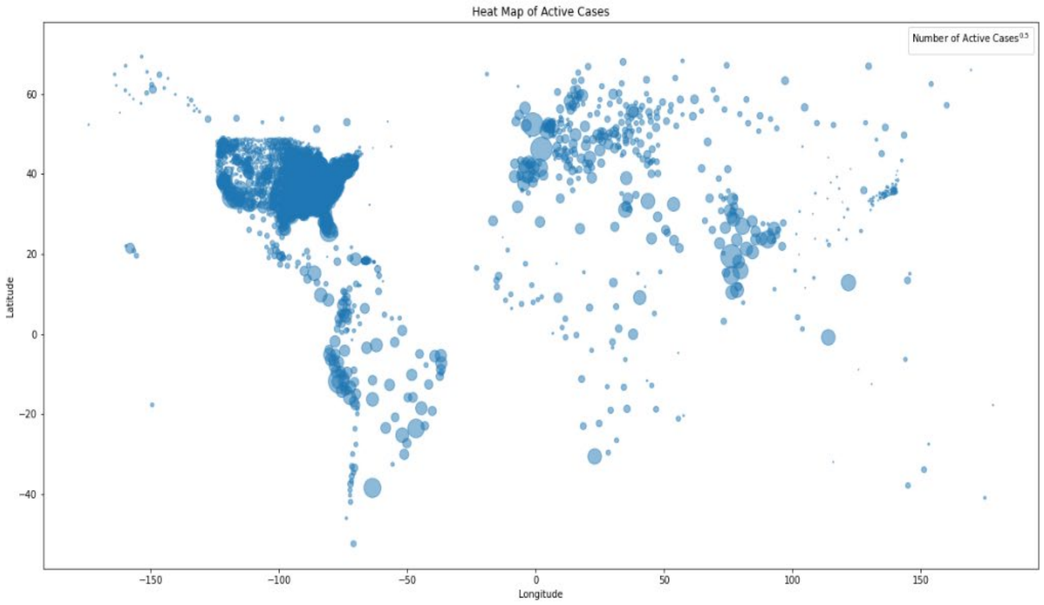

Figure 1: World Heat Map

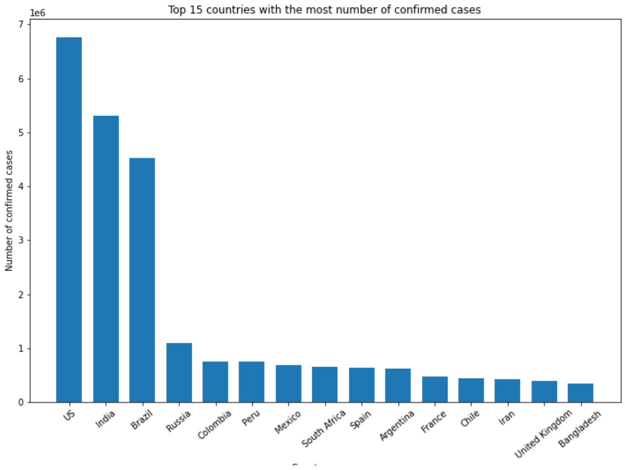

Figure 2: Top 15 countries with the greatest number of confirmed cases.

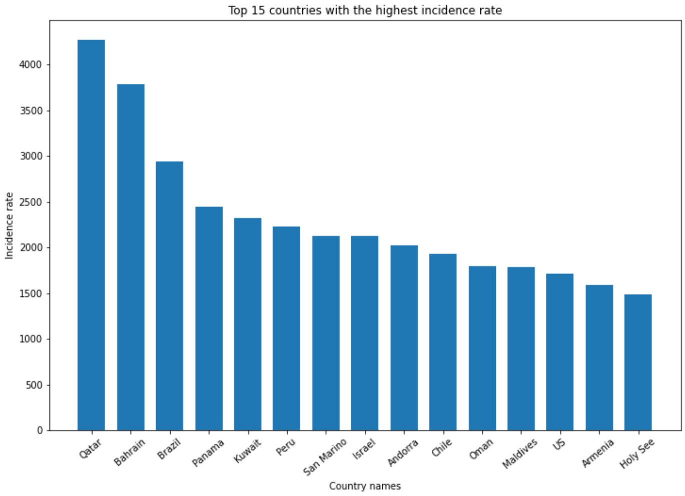

Figure 3: Top 15 countries with the highest incidence rate.

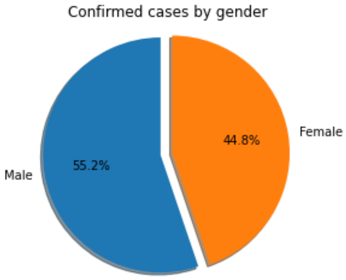

Figure 4: Confirmed cases by gender.

Before parameter tuning, we also do some visualizations to the processed data, and the results are as follows.

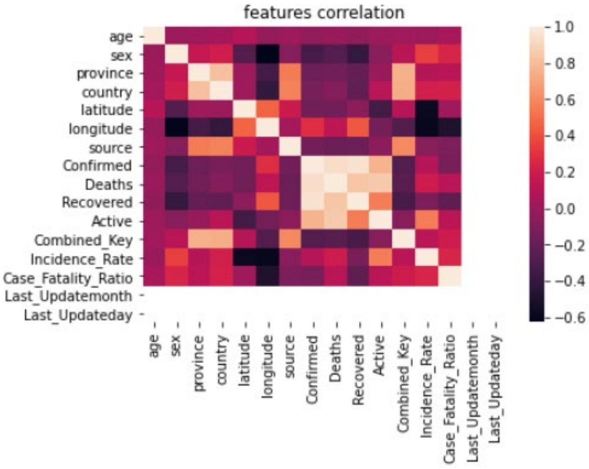

Figure 5: Features correlation.

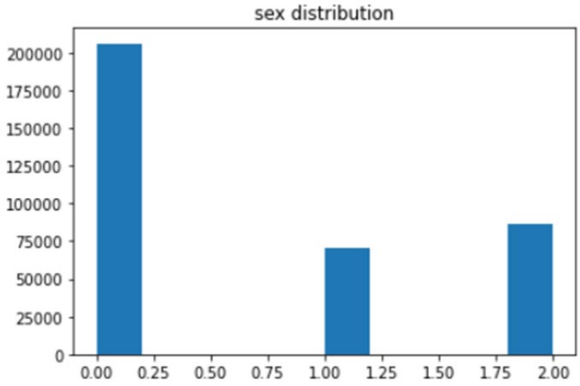

Figure 6: Sex distribution.

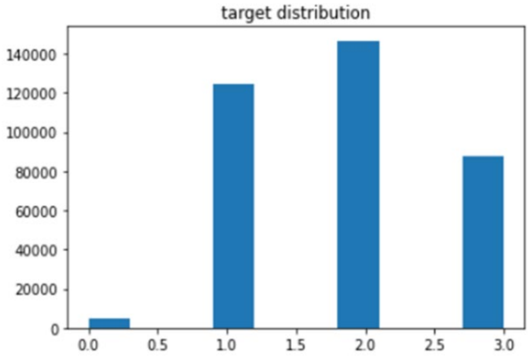

Figure 7: Target distribution.

Data Preparation

First, read the training set and test set to see if there are missing values. In some cases, fill in different types of features, such as mode for int, mean for float, etc. Check whether there are abnormal values by drawing and eliminate them if there are any.

Classification Models

Three models are chosen in this project, namely LightGBM, SVM and MLP. The three models represent the boost tree model, the linear model, and the neural network model. In this way, the experimental results can best reflect the effect of various models, to select the best model.

Reasons of choosing LightGBM: (1) Lead wise mechanism with depth limitation. Leaf wise is a more efficient strategy, each time from all the current leaves, find a leaf with the biggest splitting gain, and then split, to cycle. (2) It supports categorical features. No one hot coding is needed. In general, we need to transform category features into multidimensional one hot coding features, which reduces the efficiency of space and time. The use of category features is very common in practice. (3) Missing values do not need to be filled. (4) Sample sampling and feature sampling can be carried out to prevent over fitting.

Reasons of choosing SVM: (1) It has low generalization error rate, low computational overhead and easy to interpret results. (2) It has its own penalty term, which is not easy to over fit.

Reason of choosing MLP: It is a relatively simple neural network and choose the activation function to simulate the nonlinear relationship, hence not easy to over fit.

Choosing LightGBM, SVC and MLP is to compare the prediction effect of these different models from the perspective of three different types of models, to achieve more comprehensive model coverage and more convincing by comparing the experimental results.

Initial Evaluation and Overfitting

The results obtained for baseline models are as follows.

Accuracy of LightGBM: (running in Colab takes about 2 minutes)

```
the train acc of lgb is 0.8876399471845515

the train f1 of lgb is 0.8440141870378591

the val acc of lgb is 0.8839573075124474

the val f1 of lgb is 0.8390591915762167
```

Accuracy of SVM: (running in Colab takes about 110 minutes)

```
the train acc of svm is 0.8186119439936181

the train f1 of svm is 0.7447722632398904

the val acc of svm is 0.8178829807718758

the val f1 of svm is 0.7435654110583906
```

Accuracy of MLP: (running in Colab takes about 23 minutes)

```
the train acc of mlp is 0.85297980909416

the train f1 of mlp is 0.8088329316874047

the val acc of mlp is 0.8508926360960581

the val f1 of mlp is 0.8059151046422114
```

As shown in the figure above, the ACC and F1 scores of the training set and test set of the three models are shown. F1 score is the harmonic average of recall and accuracy, while accuracy is the accuracy of prediction. Combined with the above data, it is not difficult to find that LightGBM model has the best prediction accuracy and F1 score, close to 0.9, while SVM model has the worst prediction effect, about 0.8. It can be concluded that there is not a simple linear relationship between the objectives and features in this paper, so the prediction effect of linear model is relatively poor.

Combined with the field meaning of data, the main purpose of this task is to predict the outcome of a case, focusing on the accuracy of prediction. So, we choose to use accuracy as the most important evaluation index. If F1 is used as the evaluation index, and F1 is determined by the recall rate and accuracy, but this task does not need a high recall rate, but to predict whether the future case outcome is accurate. So obviously, accuracy is better than F1.

We also determine there is NO evidence of overfitting, because the prediction results (ACC and F1) of the training set and the test set of the model in this project are relatively close, which indicates that the model in this paper has good normalization ability, and that these models have not been fitted. To check whether it is over fitting, it is mainly to compare whether there is a big gap between the test set and the verification set. At the same time, it is necessary to check whether the parameters change, and the corresponding evaluation indexes also change significantly.

Changing LightGBM parameter learning_rate from 0.1 to 0.3. The new model is: (running in Colab takes about 2 minutes)

```
the train acc of lgb1 is 0.9001423541385856

the train f1 of lgb1 is 0.859573027731438

the val acc of lgb1 is 0.8944928891700823

the val f1 of lgb1 is 0.8516677946437203
```

The SVM kernel is changed from RBF to linear, and the degree is changed from 3 to 5. The new model is: (running in Colab takes about 113 minutes)

```
the train acc of svm1 is 0.7721646081478832

the train f1 of svm1 is 0.7368400149333164

the val acc of svm1 is 0.7704040931972602

the val f1 of svm1 is 0.7348254991977888
```

Change MLP parameter Max_ ITER is changed from 200 to 400. The new model is: (running in Colab takes about 12 minutes)

```
the train acc of mlp1 is 0.837056088906005

the train f1 of mlp1 is 0.7741578816935775

the val acc of mlp1 is 0.8366847302836079

the val f1 of mlp1 is 0.7734488285125544
```

After comparing the previous prediction indicators, we found that the indicators only change slightly. So, it can be judged that the model does not have obvious over fitting.

Hyperparameter Tuning

Parameter adjustment of gridsearchcv is what we use in this project. Take LightGBM as an example. Gridsearch needs to search the optimal value of four parameters, which are 'Max'_ depth': [4,6,8], 'num_ leaves': [20,30,40], 'learning_ rate':[0.01,0.05,0.1,0.2],'n_ estimators':[100,200,300]. But the whole search will be very slow, so we use every step to search for the optimal value of a parameter, and other parameters are fixed. This will greatly improve the efficiency of searching the optimal parameters.

Results

	Precision	Recall	F1-Score	Support
Recovered	0.10	0.87	0.18	512
Hospitalized	0.89	0.79	0.84	141380
Nonhospitalized	0.99	0.99	0.99	146524
Deceased	0.68	0.79	0.73	75114

Table 1: Results by labels.

	Precision	Recall	F1-Score	Support
Accuracy			0.87	363530
Macro Average	0.67	0.86	0.69	363530
Weighted Average	0.89	0.87	0.88	363530

Table 2: Overall results.

Conclusion

After comparing different indicators, the conclusion is that *LightGBM* model is the best in all indicators.

Prediction on Test Dataset

The same method is used to process the test data, delete the redundant irrelevant columns, and use the trained model to predict the test set.

Lessons Learnt and Future Work

During the project, we figure out how to check outliers, how to build classifiers, how to judge overfitting and how to do parameter tuning.

In the future, we can try more different models, such as complex neural network model, to do this experiment. At the same time, we will explore feature engineering to see if we can achieve better results.

References

[1] Everyday I know to learn, "Data mining Xiaobai series! LightGBM detailed explanation and tuning," CSDN, 9 July 2020. [Online]. Available: https://blog.csdn.net/qq_35679701/article/details/107239487. [Accessed 24 April 2021].

[2] i-code, "Xgboost for data mining," CSDN, 29 October 2019. [Online]. Available: https://blog.csdn.net/weixin_44132035/article/details/102807785. [Accessed 24 April 2021].

[3] swordtraveller, "[Python data mining] sklearn-SVM classification (SVC)," CSDN, 18 June 2019. [Online]. Available: https://blog.csdn.net/swordtraveller/article/details/92786837. [Accessed 24 April 2021].

[4] Shell ER, "Movie story data mining based on deep learning framework Keras and MLP model," CSDN, 13 May 2018. [Online]. Available: https://blog.csdn.net/wlx19970505/article/details/80301193. [Accessed 24 April 2021].

Contributions

The “Full Mark 459” development group presents its compliments to professor Dr. Ester and teaching assistants in the data mining course and has the honour to inform that this group consists of three students, namely Yuqing Wu (ywa292@sfu.ca), Zixi Bai (zixib@sfu.ca) and Zeyong Jin (zeyongj@sfu.ca).

During this project, each member contributes a lot. All three members collaborate to finish data visualizations and preprocessing step.

Starting from the second milestone, each group member is responsible for one classifier’s building, overfitting detection and parameter tuning. To be precise, Yuqing Wu is responsible for the LightGBM model, Zeyong Jin is responsible for the SVM model and Zixi Bai is responsible for the MLP model.

After gaining approval of all the contents in the final report from three members, Zeyong Jin writes this report.

Thank you.