Capstone Proposal

Field

Education

Background

When I was young, I heard a lot of strange questions and requests. My mom always tells me that puppy love, internet, and games would put down my grade. But could that happen? My aunt said a good relationship with the parents and even the career of the parents could affect children's grades. But, is that true? My teacher said she know how I would perform in the final exam by reading my previous tests. But, is that possible? There are so many rumors and myths in the field of education. I want to find out which are true or not.

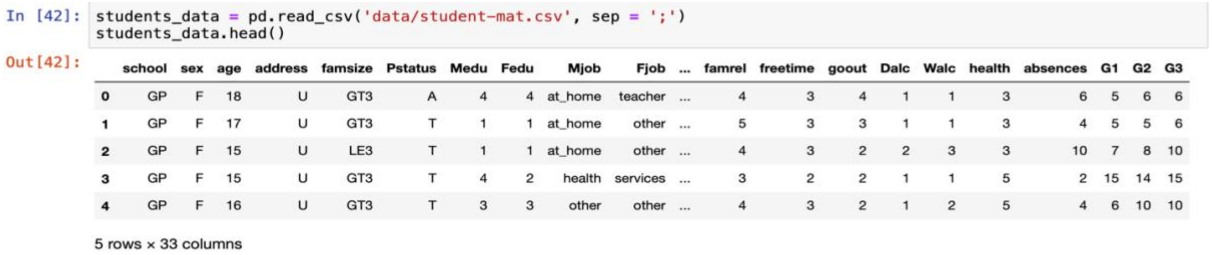
Problem Statement

I will separate the project into two parts.

1) First to find out which features (behaviors or activities) impact students' grades most. Would it be the same as rumors? Puppy love or internet could drag students' grades down to the bottom. (Multi-labels classification)
2) Second, I would try to use different datasets to forecast students' grades. For example, could we forecast the score in the next exam by using the last performance? Then what about doing the same thing by using the score before the last exam? In one word, forecast the next score by using previous behaviors and scores. (Time Series Forecasting)

Datasets & Inputs

Student Performance Data Set.
Source: https://archive.ics.uci.edu/ml/datasets/student+performance
By looking the data set, there are 33 different features in the original dataset, including 'family size', 'mom's job', 'father's job', 'health status', 'past scores' and so on. I have a lot of interesting in 'parents' status' (the relationship between parents), 'free time', 'internet', 'go out', 'relationship' (puppy love) features, because I think those columns could influence students' grades most. So I would deeply exploring those 'important' features. On the other hand, I find some features are not so important, such as 'school', 'address' and so on. Since all students in the dataset came from one school, I would drop of those columns. Original data int 'Student Performance Data Set' is object type. In this case, I would convert them into integer or float and then using one-hot encoder to extract features from the dataset.



```
In [42]: students_data = pd.read_csv('data/student-mat.csv', sep = ';')
         students_data.head()
```

Out[42]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10

5 rows × 33 columns

Solution Statement

1) Find out which features (behavior or activities) impact students' grades most. (ex. Extracting top 10 features)
2) Forecast the future exam score and compare the variance between the predictions and real score.

Benchmark Model

Relevant Paper

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Benchmark Performance

In the paper above, they achieve 60.3% accuracy in five level classification by using a neural network and the entire dataset. 30.4% accuracy in five level classification by using a neural network and the minimal dataset. ±2.05 variance in prediction score by using a neural network and the entire dataset. ±2.82 variance in prediction score by using a neural network and minus dataset. ±4.41 variance in prediction score by using a neural network and the minimal dataset. I have confidence to prove their result and get even better performance by design model and config parameters elaborately.

Evaluation Metrics

1) Because the first goal in my project is based on data mining, so I don't think there is a proper metric in this part. I could indicate a null hypothesis such as 'puppy love' and 'Internet' impact students' grades most. And set an alternative hypothesis to oppose the first opinion. After extract and calculate all features, try to accept or refuse my original hypothesis.
2) In the second part, I would set three different metrics for my predictions.
    a. Set a Pass/Fail binary classification, predict results by using different datasets. Expected performance: 90%
    b. Set a five levels multi-labels classification, predict results by using different datasets. Expected performance: 80%
    c. Predict the score in the future and compare it with the real score, calculate the variance. Expected performance: ±5.0

Project Design

Workflow

1) Explore the dataset
2) Clean the dataset, drop columns are irrelevant.
3) Preprocess the dataset, set one-hot encoder.
4) Ready to work
5) Configure model
6) Train model

7) Test model
8) Deploy model
9) Get conclusion

Algorithms

1) One-hot encoder -> preprocessing
2) Random forest -> classification
3) Recurrent neural Network -> classification
4) Recurrent network (sequential) -> time series forecasting