Qn 18) How would you build a model to predict a March Madness bracket

What kind of model?

It’s a game between two teams, and to predict who wins, we must use a classification model

The output will be binary, win or lose.

What features?

Assuming we have tons of previous data from the previous years, aside from knowing the outcomes of who wins and who doesn’t, we need to find out what kind of columns we want

Some of the features I can think off would be:

1. The amounts of shots, type of shots
2. Out of those shots, how many did it truly score?
3. Against which team
4. What are their scores? What’s their relative score from the opponent?

Rank them based on these combinations and calculate the team’s average performance across maybe the last 10 matches

How to prevent overfitting?

We will need to be cautious about putting all features together as it might like to be overfitting. One way to reduce overfitting is to remove those that are highly correlated.

For example, we do not want to include both free throws made, and free throws attempted together, as likely more throws attempted, will be more throws made.

I would create a feature that takes the percentage made/attempted, so instead of both features, I will use this for the model.

How we build the machine learning model

Since we have past data, we already know the outcome of two teams, hence we just need to train over these y variables.

Train test split, so that after training the model, can use the hold out set for testing to see if model works on unseen data.

So in the real world when using this model, we won’t have this Y variable of who wins or loses. Based on predict, we can determine who wins and who loses

How do we assess the model?

Based on its predict proba, and actual scores and results, we can do Error analysis. It could be due to change team coach that particular year etc.

QN14) How would you suggest to a franchise where to open a new store?

What kind of model?

I would use regression model, where the output will be continuous, depending on the KPI a franchise needs, let say ROI.

What features?

What does a franchise need? It needs marketability, especially important to know demographics of the place vs their financials.

Gather the whole dataset of all franchises around these locations,

1. How accessible are the places to customers?
2. Type of franchise (dessert, mains)
3. Their financials.
4. Effect of sales on weather/population density/local income levels.
5. Check of seasonality across time
6. Proximity to other sales business

We need to also understand what this franchise prioritises more?

ROI, operating profit

How we build the machine learning model

Build a model let’s say Lasso regression over Linear regression to prevent overfitting and predict better to unseen data.

We use Lasso to penalize the sum of abs values of the weights.

How do we assess the model?

We can see the importance of feature leading to the success of the franchise. Those zero-ed out means it may not be useful in predicting the success of the franchise.

We survery around different places considering franchise details, and put into our model, and let the model predict which is the best KPI score output from the regression model.

QN16) Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

Output: Donate or not (binary)

INPUT: rows (many alumnis across the years), year of graduation, major, previous donations

How we input the features

* Major, arts and science -> one hot encoding
* Year difference between graduation year and current year

Budget: income,

Columns: demographics (income, work)

Build model: Logistic regression

Assess model

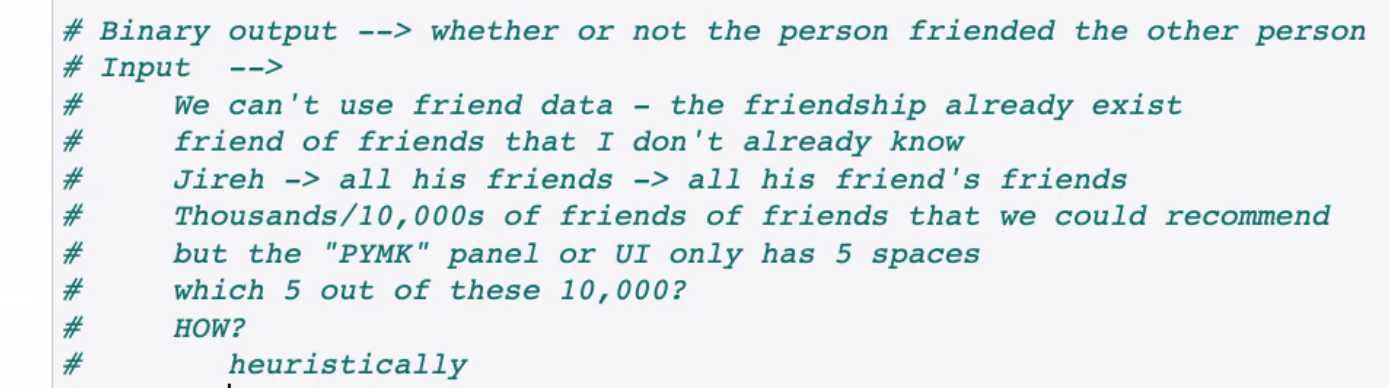
Student ID, student name, major, year of graduation , sports, honors.

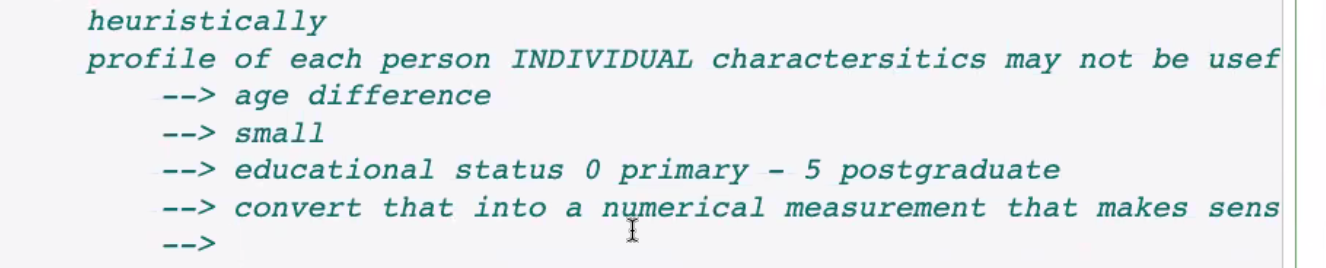
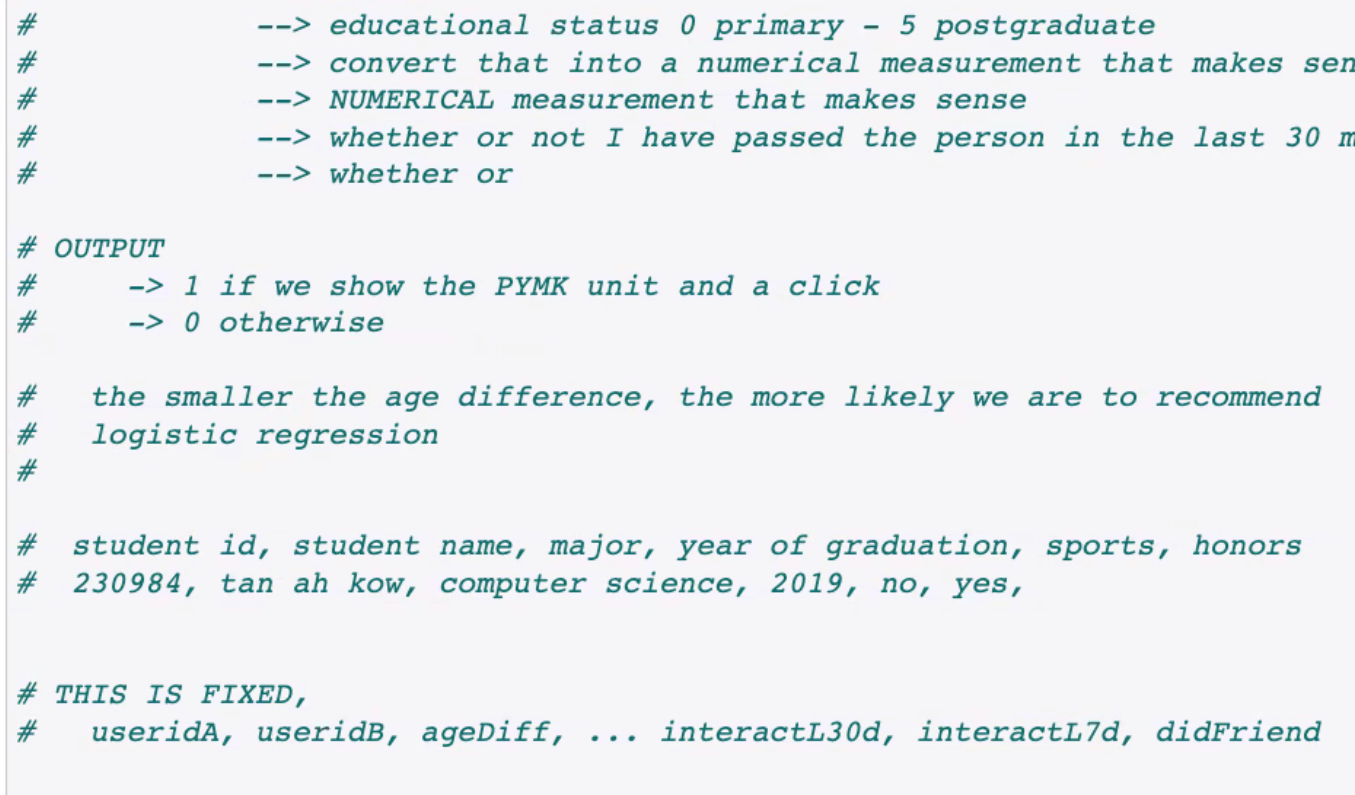
Qn 12) How would you design the people you may know feature on LinkedIn or Facebook?

Binary output – whether or not the person friended the other person

Input – we cant use friend data – friendship already exist

* Friend of friend that I don’t know



The smaller the age difference, the more likely we are to recommend

Slope of this column must be big

QN 11) How would you construct a feed to show relevant content for a site that involves user interactions with items?

