

Feedback

Group: Data Samurai

1. ACE

Do good:

Group ACE did a good job in picking the right model. They tried different models and conducted analysis on which one is better in this case. The movie score prediction part predicts some recent movies, therefore successfully attracts audience's attention. They also analyzed how to invest a profitable movie. Both sides (investors and audience) can benefit from this project.

Could do better:

As I mentioned before, this group did analysis for both sides, this further completes the analysis, but there could be a more direct logic line that connects these two parts together. Personally, I got a little confused when they transfer from movie recommendation to investors.

Improvement on analysis:

Maybe could include more visualization, especially when displaying how to invest a profitable movie.

2. Advertisement Wanted

Do good:

This topic is interesting, and they did a good job in presenting it. Besides, they clearly showed the whole process of modeling. They tried new models based on the results and weaknesses of previous models, and compare them based on comprehensive standards.

Could do better:

They could introduce the variables they used more clearly at the start of the presentation. In the presentation, they just showed a table of how they cleaned the data and directly moved to introductions of models. It took me some time to figure out what exactly the variables are.

Improvement on analysis:

When it comes to ranking the video recommendation, I think besides like/dislike count ration and sentiment score, they could also analyze the similarity of movie description and recommend movie based on that.

3. Avengers

Do good:

They asked a question at the start of presentation, which managed to arouse others' curiosity. They use different models for the analysis, and explained why they used different model, the weakness of each model, and how this leads to the next one. This shows that they know what they are doing and why they are doing this. Also, the accuracy comparison among different models is really clear.

Could do better:

In the presentation the speaker spent similar time on different parts like data processing, sentiment analysis and machine learning. They could have spent more time on the core and

explained it more clearly. They could also talk about how to generate business insights from the analysis.

Improvement on analysis:

It seems that they didn't consider the influence of different product type. Conducting analysis on different product type and exploring how they are different from each other is an interesting topic.

4. BeautifulSoup:

What impacts the airbnb spatial distribution? (where is airbnb located)

1. What affects distribution of the house

Distribution of airbnb, visualization in the first part gives a clear view of what they are doing
Location is crucial

2. Feature engineering

Correlation and importance

3. Forest, K-means

Do good:

A lot of visualization (as it is a geographical project), which is very straightforward and easily understood. Elaborate through different aspects, such as feature engineering, random forest and k-means,

Could do better:

You can try to surprise the audience by bringing out new features (rather than the ones we might know) because airbnb is a relatively "old-school" topic, and we always want to see something different.

Improvement on analysis:

You can still give a quantitative score on different important features. And trying to compare different methods using a clearer way: for example, under Random forest method, which features are the most important ones and under K-means, how will the result change. Using this way audience can have more understandings on the effect of different methods, and know which factors are the vital ones (for example, # of rooms appears in all methods, then it is quite important.)

5. DaMafia:

Crash Proof Journeys: Traffic Collisions

Do good:

It is a project that is closely related to us. Impact of the accident - introduce the accident severity so that you can understand which feature is important.

Step by step analysis so that the audience can have a whole picture of the project.

Could do better:

You can let the visualization more involved into the presentation rather than just bring them out. Like the previous group just did a great job of visualization.

Improvement on analysis:

It seems that all three model you introduce can have a better accuracy of prediction, and you can spend more than on how to improve the coefficient (like GridSearch). Time and location is important(I know that), but what is the underlying factors? It seems that all three model you introduce can have a better accuracy of prediction, and you can spend more than on how to improve the coefficient (like GridSearch). Time and location is important (I know that), but what is the underlying factors?

6. Data Massage:

PUBG

Do good:

The topic that the team chosen is very unique and is a project I am interested in.

Could do better:

Because the project is about a game, therefore the one who know it before will find it interesting but for the others who do not understand the game it might be a little bit confusing. So It could be better spending sometime on the introduction. Also, for the result they provide, they can spend some time on the connection between data analytics result and the insights in it.

Improvement on analysis:

For the analysis of game mode(solo or teamup), it seems a little bit intuitive - without the backup of analysis. (what if I choose one-person squad). How you get the data. The lading place. The visualization seems just be man-made(or copy from online). You can actually visualize based on the log and lat(which is possible in the game). Finally, for the weapon choosing analysis, It seems that the dataset is a little bit old: as far as I know, the data they used is at least 1-2 years old. It could be better to get some new data, or explain why the data is inaccessible.

7. DDOG

Do good:

The team has creatively identified the key words for later classification themselves since the impeachment for president Trump is a very new subject and there is not much research that has been done for reference.

Could do better:

There is no validation to test the model because of the same reason above. So, we wouldn't find out how good the model is.

Improvement on analysis:

They could probably use impeachments in the past in the US (through text mining on research/news in the past). For example, the most recent one was Bill Clinton in 1998.

8. Fantastic three

Do good:

The team has shown very detailed development of the film industry and its recent trends, which helped them and the audience understand what changes have occurred in this industry over time before getting into technical parts.

Could do better:

The selections for movie and film company are mostly focused in North America (United States), which might not comprehensively show the global trends and development better.

Improving on analysis:

Except for US, Australia, the film industry in United Kingdom, and probably Singapore can also be considered and compared to see the difference between different continents.

9. Hello World

Do good:

During the presentation, there was a good variety of visualisation which has shown the audience very directly about their process of data analysis and data structure.

Could do better:

However, there were too much information before applying machine learning modelling, which might not provide the best results within a limited amount of time.

Improving on analysis:

Maybe it is not necessary to include codes/data output from jupyter notebooks in the presentation, since they are not obvious enough to show the working progress. Results with some level of interpretation would be better than raw outputs.

10. Pandas

Do good:

This project has used most of data analytics tools the class has covered so far, which shows a good master of knowledge we have been learning, as well as how data can be analysed and used in a very diversified way.

Could do better:

Probably the project could be more creative rather than just doing recommendations like other groups which chose the same/similar topics.

Improving on analysis:

Different cities/countries can be analysed at the same time and the team could try to explain why the differences happen through comparison from both quantitative and qualitative perspective.

11. Suicide Squad

Do good:

The data pre-processing and regression analysis parts were well presented by both various plots on slides and verbal explanation by the team.

Could do better:

How neural network was used in the modelling was very broadly explained compared with other machine learning algorithms/modelling methods.

Improving on analysis:

There are many ways to use neural network, for example, RNN, CNN, ANN are all different ways, and the team could specify exactly which algorithm they have chosen, and if possible, try to use different ones in neural network and provide the best after comparison among all.

12. Tofu Chili:

Prediction the salary of NBA players.

Do goods:

A lot of data sources.

Good to adjust the salary to the present years (because of the inflation)

Clear visualization on the change of salary over years.

Good presentation skills

Could do better:

The connection between text mining and prediction? I might not get why there is a need for text mining, for which the team could elaborate more on it. Also, more team engagement needed - it is a team project but only one person present. Of course, he has excellent presentation skills

Improvement on analysis:

The salary that most people (fans) cares about, that are the superstar's salary, seems to follow quite different patterns comparing to the "middle class" player. These superstar salary are extremely high that could affect the league's average, however, most of the data are from those middle level players. I think it could be better to exclude the superstars (because they are unique, more factors might affect their salary such as advertisement, MVPs, Honors, even leagues' regulations like Larry Bird exception)

13. Uncle_Luoyang_Everyday:

Who dominates the box office: predict the revenue of movie.

TMDB IMDB dataset

Do good:

Models are good, clear and easily understood. We can see from the model itself which are the influential factor of the revenue. Also, the visualization provides us a better understanding of importance level for each feature.

Could do better:

It seems obvious that budge and revenue is highly correlated. And the group provide it as the final result. However, if we take movie producers' point of view, we might want to see more business implication of this project rather than some above-board facts.

Improvement on analysis:

Also, ignore the effect of "super movies" like the NBA superstars before. It could be better to include the stars and directors of the movies as a variables (or at least provide a subjective score like the DaMaFia's NY traffic collision project)

14. Untitled

Do good:

This group used different filters for recommendation, which increased the credibility of their results. They also clearly explained the challenges they met in text mining and how they solved the problems.

Could do better:

They could talk more about the application side, for example, how to make use of the results to generate values.

Improvement on analysis

The dataset they used is relatively small, this may affect the results to some extent. They could try larger dataset if possible.

15. Untouchables

Do good:

They offered a clear framework of the whole project at the very start of presentation, which makes it much easier for audience to get their points analysis. They make use of most of topics we discussed in the class, like web scrapping, visualization, machine learning etc. And the analysis is really complete.

Could do better:

Visualization part could be more closely related to other analysis. Now it seems like they just want to add visualization into the project. They didn't explain how it could provide insights the following analysis.

Improvement on analysis:

They could try more models and compare the results so as to find the one with the best performance.