Leonid Grekhov

Movie recommender report

The model I created consists of two main parts. The wide layer and a deep layer. For the wide layer I combined the genders of the users and the genre of movie they like. This created a layer that could be general enough to classify. The layer was one hot encoded to be processed.

```
df_wide
```

	F_Action	F_Adventure	F_Animation	F_Childrens	F_Comedy	F_Crime	F_Documentary	F_Drama	F_Fantasy	F_Film-Noir	...	M_Film-Noir	M_Horror	M_Musica
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
99995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
99996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
99997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
99998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.
99999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.

100000 rows × 37 columns

For the deep layer I used the age, gender, occupation and genre. These parameters are specific enough to distinguish individual users. The gender, occupation and genre columns were encoded as well as the age column was scaled.

The data was then split into an 80/20 train/test split. Based on these parameters the model predicts a rating a user would give to a genre given their age, gender, occupation and genre preference. Input layers are passed to embedding layers to be passed to a concatenate layer to be passed to several dense and dropout layers. This is then concatenated again with the wide input layer to create the final model. MSE was used

```
df_deep
```

	age	gender	occupation	genre
0	0.803030	1	15	2
1	0.803030	1	15	0
2	0.803030	1	15	5
3	0.803030	1	15	7
4	0.803030	1	15	3
...	...	...	...	...
99995	0.181818	1	18	0
99996	0.181818	1	18	0
99997	0.181818	1	18	0
99998	0.181818	1	18	3
99999	0.181818	1	18	4

100000 rows × 4 columns

for the loss and metrics while adam was used for the optimizer.

After the model finished training and evaluation over 50 epochs the loss and MSE were about 1.21 which is not that good with a max rating of 5. The model takes about 10 minutes to run.

There are several potential changes that would be made to improve the accuracy of the model. Changing the parameters for the wide portion of the model that could more precisely identify users or even combining more than two parameters. Adjusting the number of columns for the deep portion of the model. Lastly, adding more layers to deepen the model further and increase training time.