COMP4332/RMBI4310 Project 3 Report

Group 35: NANDA Nikhil (20491384), Jamie, Lo Sze Yuen (20619883), LAI, Wai Chun (20510702), LIN, Chi Wing

1. Abstract

This project aims to solve the rating prediction problem on a subset of the Yelp dataset. The data consists of users and businesses, with a rating that corresponds to the rating a user has given the respective business. In addition, various individual user attributes as well as business attributes are available to supplement these ratings. The Wide and Deep Model (WDL) has been used since it performed better than the Neural Collaborative Filtering Model (NCF) during our analysis, with a validation RMSE (Root Mean Squared Error) of 0.9996 from the former compared to RMSE of 1.0533 from the latter model.

2. Feature Engineering / Data Preprocessing

Neural Collaborative Filtering Model (NCF)

The inputs to this model are user and restaurant embeddings. No additional attributes pertaining to the user or restaurants in particular are augmented into this model. Thus, it does not involve much feature engineering.

Wide and Deep Model (WDL)

This model incorporates the additional user and restaurant specific attributes available from the Yelp dataset. There are 3 types of features constructed that are inputs to the model:

- **Continuous Features**: Both user and restaurant specific attributes are used that have continuous values. The attributes in Table 1 are included as the continuous features. Since the range of all these attributes are different from each other, they are standardized to the same range.

user_review_count	user_useful	user_funny
user_fans	user_average_stars	user_compliment_hot
user_compliment_more	user_compliment_profile	user_compliment_cute
user_compliment_list	user_compliment_note	user_compliment_plain
user_compliment_cool	user_compliment_funny	user_compliment_writer
user_compliment_photos		
item_latitude	item_longitude	item_stars
item_review_count	item_is_open	

Table 1: Continuous Features used in the WDL model

- **Deep Categorical Features**: The 3 attributes used are "item_city", "item_postal_code", "item_state". These 3 attributes correspond to the restaurant and are used as embeddings.

- **Wide Features**: Each business also corresponds to a variety of categories. For example, “Bars” and “Nightlife” are a common pairing of categories for many businesses. Thus, in addition to individual categories, a cross product transformation of these categories is also taken consisting of 2, 3 and 4 categories respectively. The 50 most common 2 category combinations are shortlisted as well as the 30 most common 3 category and the 20 most common 4 category cross products. Finally all these categories are combined and created into individual binary features (0/1 represents whether the restaurant belongs to the respective category / cross product of categories)

3. Implementation

Neural Collaborative Filtering Model (NCF)

The user and business embeddings are concatenated before they are passed through a Multi Layer Perceptron. In addition, Dropout layers have been used to prevent overfitting. Embedding size of 10 performs the best with RMSE of 1.0533. Thus, the NCF model is a great model to get an initial estimate of the toughness of the problem as it is straightforward to implement and gives decent results on this dataset.

Wide and Deep Model (WDL)

There are two parts to this model. The wide part which corresponds to the memorization of frequent co-occurrences of business categories. Thus, it essentially serves as a linear model of raw categorical features. The deep part of the model concatenates the continuous features as well as the deep categorical features and then they are passed through a single dense layer and then a dropout layer. The objective of the deep part of the model is to generalize well based on the various inputs it has been passed to learn from.

The outputs of both the wide and the deep part are concatenated and then passed through a Dense layer to predict the final rating. The Wide and Deep Model resulted in an RMSE of 0.9996 which is an improvement compared to the NCF model.

4. Evaluation

Model	Train RMSE	Validation RMSE
NCF (embed_size=10, epochs=5)	0.86	1.05
NCF (embed_size=50, epochs=5)	0.81	1.07
NCF (embed_size=100, epochs=5)	0.78	1.06
WDL (1 dense layer for deep model)	**0.98**	**1.00**
WDL (3 dense layers for deep model)	0.99	1.01

Table 2: NCF and WDL model comparison

It can be seen that the NCF models overfit more in comparison to the WDL models. Smaller embedding sizes prevent the overfitting of the NCF model, whereas, fewer dense layers prevents overfitting of the WDL model. Thus, it can be concluded that the additional information on the users and the businesses indeed help make better predictions.