PREDICTION OF THE MEAN RATING FOR A MOVIE

DATA ANALYTICS PROJECT PRESENTATION

Michelle Zanotti [1038859]
Antonio Iannotta [1024859]

In this presentation:

1. Introduction
2. Methodology
    2.1. Traditional non-deep supervised ML techniques
    2.2. Supervised ML techniques based on neural networks
    2.3. Supervised ML technique with deep models for TabularData
3. Implementation
    3.1. Traditional non-deep supervised ML techniques
    3.2. Supervised ML techniques based on neural networks
    3.3. TabNet
4. Results
    4.1. Traditional non-deep supervised ML techniques
    4.2. Supervised ML techniques based on neural networks
    4.3. TabNet

INTRODUCTION

Introduction

GOAL: prediction of the mean rating for a certain film starting from its characteristics.

DATA: MovieLens 25M dataset, composed of several files in which are stored the characteristics of the movies.

FILES

- tags.csv : tags that can be assigned by a specific user.
- movies.csv : relevant informations about every movie (movieId for a film, the title of the film and the genre or the several genres)
- genome tags.csv : list of the possible assignable tags
- genome scores.csv : relevance of every tag for every film.
- ratings.csv : list of ratings for the movies.

METHODOLOGY

Methodology

To reach the goal of the analysis → use of various techniques and approaches based on regression

Three different independent tasks but with the same goal:

- ❏ Traditional non-deep supervised ML techniques
- ❏ Supervised ML techniques based on neural networks
- ❏ Supervised ML technique with deep models for Tabular Data

Methodology

The methodological approach followed during the development of the various tasks followed the several steps of data analytics pipeline:

- ❑ Data acquisition
- ❑ Data visualization
- ❑ Data pre-processing
- ❑ Modelling
- ❑ Performance evaluation

Methodology: Traditional non-deep supervised ML techniques

Linear Regression

basis of the regression

to obtain the estimate of a function a parametric approach is used by calculating the parameters (coefficients) that constitute it

KNN Regressor

use of a method based on the calculation of distances

(adaptation of the knn classifier)

Random Forest Regressor

being a tree-based assembly method it should increase performance even if overfitting is more likely

Methodology: Supervised ML techniques based on neural networks

To develop an approach based on neural networks, three fundamental elements were necessary:

1. the creation of a data layer
2. the architecture of the network → deep feedforward network (multi layer perceptions)
3. training and the evaluation process

Methodology: Supervised ML technique with deep models for Tabular Data

Supervised ML technique with deep models for Tabular Data → TabNet.

high-performance and interpretable canonical deep tabular data learning architecture

canonical DNN architecture for tabular data

Methodology: Supervised ML technique with deep models for Tabular Data

Main features

- Usage of sequential attention to choose which features to reason from at each decision step
- the learning capacity is used for the most salient features → interpretation and learning are more efficient
- Receiving raw tabular data inputs without any preprocessing and is trained using gradient descent-based optimization.
- Two kinds of interpretability:
  - local interpretability that visualizes the importance of features and how they are combined
  - global interpretability which quantifies the contribution of each feature to the trained model

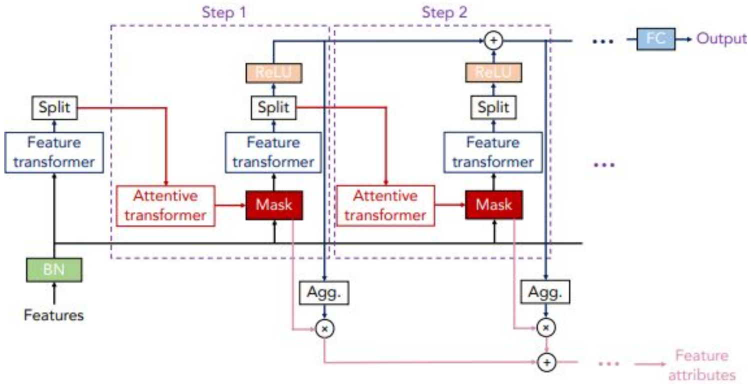
Methodology: Supervised ML technique with deep models for Tabular Data

The TabNet encoder

sequential multi-step, where inputs go from step to step

A step is composed of:
- ❏ feature transformer
- ❏ attentive transformer
- ❏ feature masking

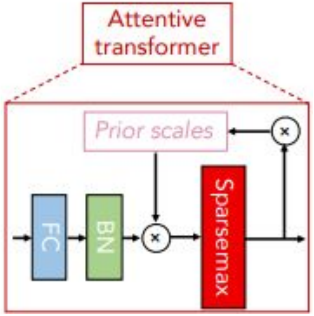
A split block divides the processed representation to be used by the attentive transformer of the subsequent step as well as for the overall output.



For each step:
- ❏ the feature selection mask provides interpretable information about the model's functionality
- ❏ the masks can be aggregated to obtain global feature important attribution.

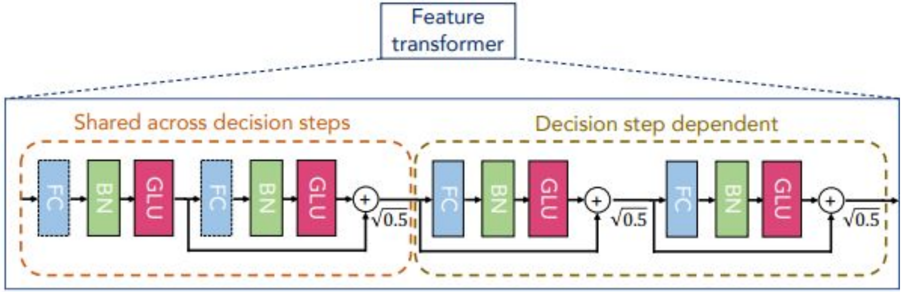
Methodology: Supervised ML technique with deep models for Tabular Data

Attentive Transformer



Feature Transformer

Normalization with 0.5 helps to stabilize learning by ensuring that the variance throughout the network does not change dramatically.



Methodology: Supervised ML technique with deep models for Tabular Data

Feature selection
Carry out a soft selection of the salient features by creating a mask (M[i]) using an attentive transformer that allows to obtain the masks using the processed features from the preceding step;

Feature processing

Processing of the filtered features using a feature transformer

Division by the output of the decision phase (d[i]) and the information for the next phase (a[i])

Aggregation
The outputs of each decision step are combined in a linear way.
It quantifies aggregate feature importance in addition to analysis of each step

Combining the masks at different steps → a coefficient that can weigh the relative importance of each step in the decision

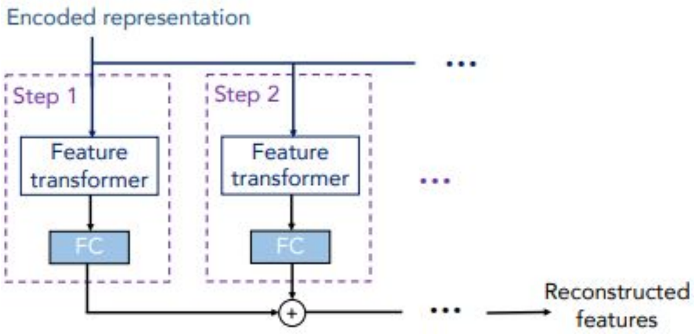
ηb[i]:

- used to scale the decision mask at each decision step obtaining the aggregate feature importance mask.

Methodology: Supervised ML technique with deep models for Tabular Data

The TabNet decoder

used to perform Tabular self-supervised learning by re-constructing the tabular features from the TabNet encoded representation.



It is composed of

- ❏ feature transformers
- ❏ fully-connected layers at each decision step

The outputs are summarized to obtain the reconstructed features.

<u>What is done</u>

the prediction of the missing columns from those present by exploiting the binary mask S multiplied with the last FC layer

the input coming from the encoder is given by $(1 – S) \cdot \hat{f}$

IMPLEMENTATION

Implementation: Traditional non-deep supervised ML techniques

Three different models trained in this task:

- Linear regression
- KNRegressor
- RandomForest Regressor

The several steps applied to each model are the following:

- Scaling
- PCA (70%)
- Hyperparameter Tuning
- Evaluation of performance

Implementation: Supervised ML techniques based on neural networks

For this task have been performed:

- definition of architecture
- definition of data layer
- definition of train and test model

The execution, in order to evaluate the performance, has been performed with dropout and without dropout, in order to make a comparison.

Implementation: TabNet

The last task required to use the tabular NN. In this step a Tabular neural network has been created starting from the TabNet model and just changing some parameters (like the one to auto select the best learning rate).

RESULTS

Results: Traditional non-deep supervised ML techniques

In the following are reported the results for the first task

Algorithm	MSE	R^2
KNN regressor	0.04313929340398458	0.8127800038849485
Linear regression	0.010761137943908726	0.9532977931468383
RandomForest Regressor	0.08046306577307188	0.6507987574027296

Results: Supervised ML techniques based on neural networks

In the following are reported the results for this task

Dropout	Loss	R2
No	0.03749039024114609	0.84
Yes	0.06291425973176956	0.63

Results: TabNet

In the following is reported the results for the best TabNet model

MAE	MSE	R2
0.1651230036436534	0.043966363347998964	0.803270316811896