Task 1 | Supervised Learning | AI & Computational MIS

Prediction

Prediction-oriented modeling plays an important role in nowadays practice: the aim is to predict the as yet unknown target variable *y* for given, new observations, e.g., to predict the housing prices - as accurately as possible - for new houses against the background of available housing data. This is done on the basis of the existing data of the prices from the housing market, i.e., including the selling prices known for these houses (supervised learning).

A distinction is made between two sets of the partial data: On the one hand, there is training data (also called learning data), which originates from a learning or estimation sample, and on the other hand, there is testing data, to which one applies the model.

1. In the training data, both the explanatory variables x = (x_1, x_2, ..., x_n) and the target variable y are present. On this training data, the model y = f(x) is formed and estimated by $\hat{f}$(x).
2. This estimated model $\hat{f}$(x) is applied to the application data x, for which (initially) the target variable y is unknown, i.e., $\hat{y}_0$:= $\hat{f}$(x_0) is computed. The unknown value y_0 of the target variable y is predicted by $\hat{y}_0$.

Possibly at an even later time the true value y_0 of the target variable y is available. Then, the own prediction $\hat{y}_0$ can be evaluated, i.e., the error y_0 - $\hat{y}_0$ between the predicted value $\hat{y}_0$ and the true value y_0 can be analyzed.

In practical applications, three successive phases can be distinguished in time (compare above):

1. The training phase, i.e., the phase in which both explanatory (x) and explained variable (y) are known. Here the model is estimated (i.e., learned): $\hat{f}$(x).
2. In the following application phase, only the explanatory variables (x) are known, not y_0. Based on the results from phase 1.$\hat{y}_0$:= $\hat{f}(x_0)$ is predicted.
3. After that a potential evaluation phase follows, for which the target variable (y_0) is known. This allows to check the prediction quality of the model.

On a computer, we can simulate this application scenario: we randomly divide the data set into a learning or training sample (training data; (x, y)) and a test sample (application data, x_0)): Modeling is done on the training data. The model is applied to the test data (application data). However, since the target variable (y_0) is also known here, the model can be evaluated with it.

Prediction Quality

Your task: Play the data scientist. Construct a model based on the training data (x, y) and predict the target variable ($\hat{y}_0$) as accurately as possible for the application data (x_0).

Your instructor(s) knows the value of the target variable (y_0). For the evaluation of the prediction quality, we use the root mean square error (RMSE) on the application data:

$$RMSE_{test} = \frac{1}{n_{test}} \sqrt{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}$$

Where y_1 are the true values, $\hat{y}_1$ are the predicted values of the estimated model $\hat{f}$(x), and n_{test} is the number of observations of the test data set (application data set). Therefore, for a good prediction RMSE should be as small as possible.

Hints

- It is up to you which variables you use for modeling - and whether you possibly preprocess them, i.e., transform, summarize, clean up outliers or similar. Just remember to perform the data transformation that you perform on the training data also on the test data (application data) as well.
- Consider criteria for model and/or variable selection. There are algorithms and Python functions for this as well.
- Avoid over-fitting
- You have free choice of methods for modeling and preprocessing: You can, for example, calculate a linear regression with variables of your choice; you can also use tree methods or apply neural networks.
- **Everything you do, data pre-processing, modeling and applying, must be transparent and reproducible.**

Evaluation Criteria

- **Formalia**: among others, reproducibility of the analysis, readability of the syntax, clarity of the analysis, comprehensibility
- **Method**: e.g., methodical correctness in the explorative data analysis, data preprocessing, variable selection and modeling method, parsimony
- **Content**: among others, correctness of the description and the interpretation of the prediction quality, good visualizations.
- **Prediction quality**: The prediction quality of the zero model corresponds to a 4.0, that of an (unknown) simple reference model of your instructor to a 2.0. Your evaluation is done according to your prediction quality, i.e., if you are better than the reference model, you will receive a better grade than 2.0 in this aspect.
- **Uniqueness:** The quantitative data analysis in execution and interpretation is the main focus of this work. Identical procedures, e.g., in the Python/Jupyter Notebook code, are by chance very unlikely and can be evaluated as plagiarism.
- **If you are hypothesis-driven:** Make sure that the null and alternative hypotheses are formulated correctly, and that the test result is interpreted correctly.
- **The overall score** does not have to be the arithmetic mean of the sub-scores. Individual particularly good or weak aspects in the sub-scores can influence the overall grade upwards or downwards.

Data Description

The dataset examines predictors of housing prices. It is not a study with explanatory or causal claims, but the quality (accuracy) of the prediction of price is the focus.

Target variable: price

Predictor Variables:

- squareMeters - construction side size in sq m
- numberOfRooms - number of rooms
- hasYard - value that tells if the house has a yard
- hasPool - value that tells if the house has a pool
- floors - number of floors
- cityCode - zip code
- cityPartRange - the higher the range, the more exclusive the neighbourhood is
- numPrevOwners - number of previous owners
- made - year of construction or renovation
- isNewBuilt - value that tells if the house is newly built or renovated
- hasStormProtector - value that tells if the house has a storm protector
- basement - basement square meters
- attic - attic size in sq m
- garage - garage size in sq m
- hasStorageRoom - value that tells if the house has a storage room

- hasGuestRoom - number of guest rooms

The dataset “train.csv” contains the target variable (y, pay), based on this data you can develop ("train") your model, it will be tested with the test data set “test.csv”. This does not contain the target variable. The distribution was done randomly. Create a model based on the observations to predict the target variable. Apply your model to the observations and thus predict the price for these observations.

Please prepare a video presentation of 15 minutes upload the files in STUD.IP until 23:59 16.01.2024