WIZELINE

ML Modeling Challenge

Data Science Hiring

Challenge

You are presented with a multivariate **regression** problem and a blind test.

Data description

In the file *training_data.csv,* you will find the training data set of **800 samples**. Each sample has **20 features**, named *[feature_0, feature_1, … , feature_19].*

As a data scientist, you will be in charge of building and training a model to predict a target variable that you will find in the *target* column of the *training_data* file.

You are also asked to provide predictions on a test set of data. The file *blind_test_data.csv* contains **200 samples** with the same 20 features as above; you must predict the missing *target* value for these samples using the model you created using the training data.

Tasks

Using the programming language and libraries of your choice, your tasks are the following:

1. Preprocess the features if necessary (justify if not).
2. Select a subset of features (justify if not).
3. Train a model using the training data set.
4. Perform the model metrics that you consider necessary or best to evaluate the performance of the model you just trained. Beware of overfitting. The target has some noise, even if you had the exact noiseless function you would get around 0.92 R^2.
5. Predict the target values with your model for the blind test dataset.

Submission

You will submit by email to **ds_interviewers@wizeline.com**:

1. Your code for the challenge, with brief comments explaining its purpose.
2. Your 200 predictions for the blind test in csv format with a single column ***target_pred***.