Crab Age Prediction

Abstract

On this project I predict crab's age from the dataset as it is one of major food in lots of countries and I would like to know how this species stands in ecosystem to protect if necessary for future generations. Crabs make up 20% of all marine crustaceans caught, and around 1.5 million tonnes are consumed worldwide annually. They are prepared and eaten as a dish in many different ways all over the world. Crabs start to reproducing when they are between 12 – 18 months, female release the newly hatched larvae between 1 to 4 million into the water during the spring and early summer, where they become part of the plankton. In our dataset we have 10 columns and more than 74k rows, columns are sex, length, diameter, height, weight, shucked weight, viscera weight, shell weight and age. Dataset is a collection of various crabs from various regions.

Related Work

For this project to predict crab's age I have used 3 different methods, each provide similar accuracy but TensorFlow Sequential model with 'Adam' optimizer looks best. There were some data cleaning involved and normalization of values for better model training. Predictions are at 57.8% accurate which provide some value, could be better if dataset will consist just only type of crabs and one region. If I can find better dataset I will probably reuse this methods.

Dataset and Choice of Methods

For this project I have used dataset from [1] Kaggle website. It consist 74k records and 10 columns with various values like sex, length, diameter, height, weight, shucked weight, viscera weight, shell weight and age. Problem with this dataset is that consist a collection of crabs from various regions of the world and various types of crabs, it is making our prediction less accurate as perfect dataset should be from one crab species and from one region, than we should get most accurate predictions for that region.

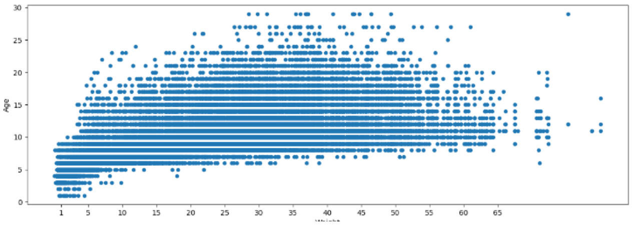
For this project I have used 3 methods to make predictions. First [2]Random Forest Regression Model from scikit-learn - a random forest is a meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Second [3]TensorFlow Sequential groups a linear stack of layers into a Model with 'Adam' optimizer. Third [4] Linear Regression Model from scikit-learn - is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables.

Methodology and Pre-processing

I have used CRISP-DM to achieve my project results. This methodology have 6 steps: 1. Business Understanding which include; defining business objectives, assessing the current situation, determining data mining goals, producing a project plan; 2. Data Understanding which include; gathering initial data, describing data, exploring data, verifying data quality; 3. Data Preparation which include; selecting data, cleaning data, constructing data, integrating data, formatting data; 4. Modelling which include; selecting modelling techniques, designing tests, building the model, assessing model; 5. Evaluation which include; evaluating results, reviewing the process, determines the next step; 6. Deployment which include; planning deployment, monitoring and maintenance, reviewing the project, finalizing the project. I have define business objectives to be - predicting crabs age. At the current point I decided to find dataset with at least 10k records and min 10 columns of values. I have planned to find dataset than clean and prepare data for use with multiple models, than evaluate results. I have got dataset from Kaggle which had all my requirements. Dataset consist 10 columns and more than 74k rows, columns are sex, length, diameter, height, weight, shucked weight, viscera weight, shell weight and age. For data preparation I did few things: 1. Load data as pandas data frame, 2. Selected all values to be used by model, 3. Checking for missing values(no missing values in each column), 4. Creating 2 separated pandas data frames, one X with all values except 'Age' column and second data frame Y which include just column 'Age', 5. Removing from X data frame column 'Id', 6. As column 'Sex' in X data frame have no numerical values I have used get_dummies function from pandas to create 3 additional columns: ' Sex_F', ' Sex_I', ' Sex_M' with numerical values 0 and 1, 7. I have used Standard Scaler from sklearn to normalize values in X data frame, 8. For splitting data to train and test data I have used train_test_split function from sklearn.

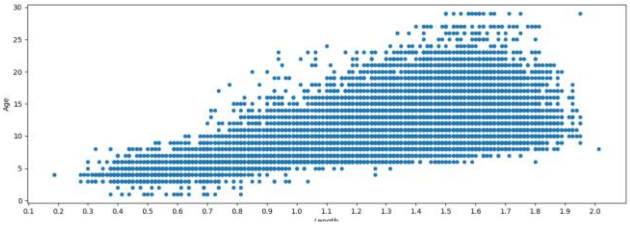
Evaluation and Presentation

Dataset that was provide include various crabs from different regions and different types of crabs. They metrics are varied that way models can't get predictions at high level of accuracy. It could be much better if we could get dataset for one type of crab and from this same region, then we if we train our models from this dataset, models predictions should be much more higher than is now. Like on picture 1 which is showing 'Age' and 'Weight' of crabs, we see that crabs have high weight for many various age.



Picture 1.

Also on picture 2 which is showing 'Age' and 'Length' of crabs, we see many of crabs that their length is huge for various age.

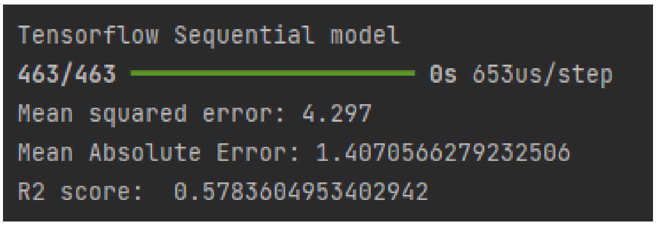


Picture 2

This all make our model confused from this dataset.

To measure accuracy I have used 3 different methods: [4]Mean Squared Error(MSE) it is used to asses quality of predictor it is the average squared difference between the value observed in a statistical study and the values predicted from a model, [5]Mean Absolute Error(MAE) in statistics it is a measure of errors between paired observations expressing the same phenomenon, and [6]R2 an R-Squared value shows how well the model predicts the outcome of the dependent variable.

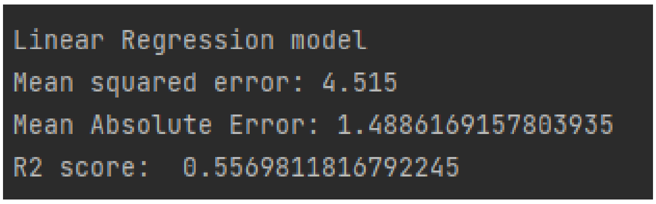
On picture 3 we see TensorFlow model accuracy:



```
Tensorflow Sequential model
463/463 ━━━━━━━━━━━━━━━━━━━━ 0s 653us/step
Mean squared error: 4.297
Mean Absolute Error: 1.4070566279232506
R2 score:  0.5783604953402942
```

Picture 3

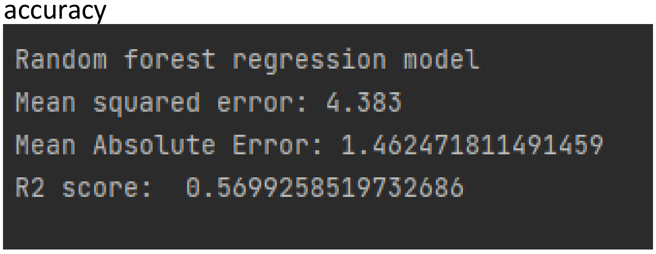
On Picture 4 we see Linear Regression model accuracy:



```
Linear Regression model
Mean squared error: 4.515
Mean Absolute Error: 1.4886169157803935
R2 score:  0.5569811816792245
```

Picture 4

On Picture 5 we see Random Forest Regression model accuracy



```
Random forest regression model
Mean squared error: 4.383
Mean Absolute Error: 1.462471811491459
R2 score:  0.5699258519732686
```

Picture 5

In MSE we want value to be as small as possible and we see that TensorFlow model have the lowest which mean have best predictions according to this measure. In MAE we want value to be small as possible also and we see that TensorFlow model have best result in predictions if we consider this measure. For R2 score we want value to be closer to 1 and we see that TensorFlow have the highest number, it mean have best predictions if we consider this measure. Like we see in all measure we took TensorFlow comes out the best in predicting age of crabs.

Conclusion and future work

I found that this prediction methods works well and could work better if I could get better dataset that is not mixture of various types of crabs and from various regions but one type of crab and one region. If I could have more time I will look for better dataset.

Video presentation link

https://youtu.be/RBGpkUFWLw4

References

[1] https://www.kaggle.com/competitions/playground-series-s3e16/data?select=train.csv

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[3] https://www.tensorflow.org/api_docs/python/tf/keras/Sequential

[4] https://www.britannica.com/science/mean-squared-error

[5] https://en.wikipedia.org/wiki/Mean_absolute_error

[6] https://www.freecodecamp.org/news/what-is-r-squared-r2-value-meaning-and-definition/#:~:text=R%2DSquared%20values%20range%20from,50%25%2C%20and%20so%20on.