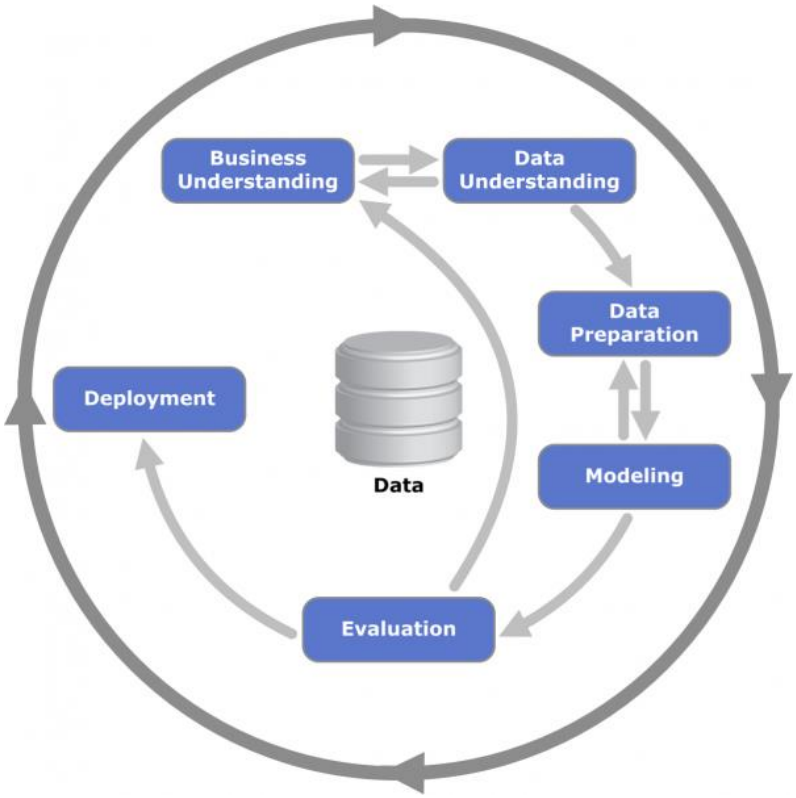
B9DA109 Machine Learning: CA_ONE

Supervised Machine Learning – Regression

November 2022

Submitted by:
Vaibhav Tiwari: 10621051
Lecturer: Courtney Ford

As a part of the continuous Assessment – 1, the team of three subdivided and achieved the stages together. Our first aim was to analyze and divide the project into modules, I suggested that we use and follow the CRISP DM methodology as suggested by our statistics professor in his initial lectures. This methodology helped us understand the steps required and what was expected in general from a project based on models based on datasets.



The prime challenge of this project was to find a suitable and fitting dataset. I went across multiple domains on the web and provided my inputs to the team. The team then came back with all of the individual inputs and picked a dataset which fit the interest of all the team members and also seemed to be the most suitable match for the expected use case indicated by the professor. The dataset picked was FIFA (Football game-based data) dataset from Kaggle. Starting with initial data collection, the I proceeded with activities to get familiar with the data, identify data quality problems and discover first insights into the data. In this phase, I could conclude that we needed to include 15 of the 60+ attribute columns that associated with the filter we wanted to achieve. The independent variables selected were purely based on the filter applied initially that the players to be analyzed would be strikers.

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Our team evaluated, selected and applied the appropriate modelling techniques.

“Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data”. I helped the team to clean and prepare the data in order to maximize efficiency while processing and computation. We performed various data checks to ensure singularity and also removed currency symbols and changing symbolic values like 10K to 10000. In this process I recommended that we ensured we converted the exponential values as well to avoid calculation errors in later stages.

Variable/Feature selection from all the selected columns was of utmost importance as the dependent variable would be computed based on these selected independent variables. In this project, I understood and applied the machine learning model to players that had been listed in the data as a Striker.

Post completion of the cleaning and preparation relevant steps, I had a dataset in-hand that was ready to build a regression fit model upon. In the initial stage I had considered single linear regression model with only a single independent variable. But, after rigorous discussions with the team we added multiple additional feature variables that seemed just in order to improve the model. Thus, we finalized multiple independent variables and one single dependent variable which led me to build a multiple linear regression model. Once built and on successfully running the model, it proved to be the appropriate model without the use of any optimization algorithm. This in turn lead us to implementing the optimization algorithm to compare the predictions. I chose gradient descent optimization algorithm to achieve more accurate predictions.

Multiple Linear Regression is a model where we can train the data and predict the output of the dependent variable. During this process, I got to know and use sklearn library and implement LinearRegression object to predict the output of the dependent variable.
LinearRegression function is used to get Linear Regression object and is used to train the data using fit method and predict the dependent variable using the predict method.

Gradient Descent optimization algorithm also known as steepest descent, improves the cost function by finding the base point using multiple iterations. This base point is also referred as global minima.

These stated details and information helped me to understand and execute the algorithm and model.

CONCLUSION:

The key takeaway from the project implementation included us being able to understand the stages involved in data selection and in turn to understand the various steps required in performing machine learning. To handpick a dataset that fits the model and aligns with the idea of group to have been able to perform and execute machine learning. The data needs to be cleaned and prepared to achieve accuracy and to avoid inconsistency and rework. Model development is key stage in machine learning for choosing which models and algorithms fits best based on the dataset. In this project, we have used Multiple Linear Regression model and Gradient Descent algorithm to perform predictions about players potential based on the selected independent variables. The cost function obtained using Gradient Descent algorithm proves to be accurate as compared to Multiple Linear Regression. After doing a comparative study / research Gradient Descent algorithm gains better accuracy over Multiple Linear Regression model.