Machine Learning Engineer Nanodegree

Capstone Proposal

Nikolas Mesquita
MAY 1st, 2019

Proposal

Domain Background

Electronic games move the billionaire market reaching 90 billion dollars in 2018 (Source). Industry giants like Sony, Nintendo, Microsoft and now Google make several monthly releases and of course the vast majority of these launches use at least one engine that employs the use of artificial intelligence techniques.

A very successful video game is FIFA which is a very popular soccer simulator and sales leader over the years.

FIFA Ultimate Team is a modality within the game where there are cards with the individual skills of the players who add up their overall maximum. When you receive these cards they are either won or bought with real or virtual money, you can set up your team (as if it were a coach) and play games that receive more cards or virtual cash prizes.

As electronic games are my favorite hobby and game since my 6 years of age, both in the virtual world and in the real world I am a soccer fan so I thought about doing this study (avoiding one of the problems of the corporate world which I work daily).

There are some studies about this background as follow:

Analyzing and Predicting European Soccer Match Outcomes
The ultimate Soccer database for data analysis and machine learning
Fifa 19 (Machine Learning)

Problem Statement

The basic idea is to analyze the set of skills and other player's information in order to predict their "overall" feature value. As this feature has a discrete variable characteristics (amounts or values) we have a Regression Predictive Modeling problem.

Datasets and Inputs

Each row on the dataset represents a football player (card) that can be used in game. Each single player might have multiple versions of himself (upgrades, transfers or special versions) and each column contains player's informations.

The data set contains 18 thousand players with 95 different attributes, among then:

- Quality of the player card: Special, Gold, Silver or Bronze
- Rarity: Rare or Normal
- All the player skills such as: pace, pass, defense...
- Player overall
- Position in the pitch and so on…

For the supervised model development, i considered the FIFA 19 data set available from https://futbin.com. All this data set attributes may help to determinate the player overall.

Solution Statement

The solution will be a supervised model capable of predicting the overall of a player given its attributes. I'll use Pandas and Numpy to gain some understanding of the data, so i will try to get some new features based on the given features in order to train the model. For the model i will try several supervised models, in which i will choose the best for the problem.

In this way, I will do an "overall" player's feature analysis, where I will randomly remove some "overall" data from the original data set. These indices will serve as a benchmark for model performance when compared to the predicted outcome.

Benchmark Model

As i chose a supervised learning algorithm to solve the problem, for benchmark i will use a naive bayes predictor, and i will try to beat this benchmark model with my solution. So i'll train and test this benchmark model using the exact same data and conditions that i'll use for my solution.

Evaluation Metrics

As the original data has the "overall" feature already filled and i will randomly removed it from the data set, after the "overall" feature prediction the evaluation metric will be only to compare the model result with the original data from data set. The metric will be how the prediction of the feature "overall" will be closer of the original information. For this i plan to use this measurements:

Mean Absolute Error(MAE) is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. It gives an idea of the magnitude of the error, but nothing about value over or under predicting.

Mean Squared Error(MSE) pretty similar MAE in that it provides only a gross idea of error magnitude. The difference lies on the conversion of units back to the original for the output variable so it can be more meaningful for data visualization. This measure is called the Root Mean Squared Error (RMSE).

R^2 (R Squared): this measure provides an indication of how good was the fit of a set of predictions to the actual values. This is a value between 0 and 1 for no-fit and perfect fit respectively.

Reference: [Metrics To Evaluate Machine Learning Algorithms in Python]

Project Design

- Programming language: Python 3.6
- Library: Pandas, Numpy, Scikit-learn, Matplotlib
- Algorithms (Support Vector Machine, Linear Regression, Decision Trees, Neural Networks and so on)

- Workflow:
    - Understanding of the dataset: create basic statistics.
    - Pre-processing: perform cleaning, encoding categorical data, handle with outliers and missing values. May be will be necessary apply some technique to fulfill this data an also run normalization and standardization.
    - Feature engineering: Create new features based on the original features on the data set.
    - Make some observations: what's mean age of the players, what the preferred foot, is there some preferred position for the players? how often a player appears as a special card, etc.
    - Fine tune the model hyperparameters. Here i can use randomize search, where there's no need to try out all parameters, or a Grid Search with an exhaustive search over a specific parameter value.
    - Perform a benchmarking model run.
    - Perform supervisioned model training on the given data to achieve the prediction
    - Run the algorithms predictions.
    - Compare the algorithms predictions results.
