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)

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.

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, we can see the advantages / disadvantages of each one (Reference).

Random Forest

In Banking it is used to detect customers who will use the bank's services more frequently than others and repay their debt in time. In this domain it is also used to detect fraud customers who want to scam the bank. Considered as a very handy and easy to use algorithm, can be used for both regression and classification tasks and that it's easy to view the relative importance it assigns to the input features. Overfitting won't happen so easy to a random forest classifier. A Large number of trees can make the algorithm to slow and ineffective for real-time predictions. these algorithms are fast to train, but quite slow to create predictions. Accurate predictions requires more trees, which results in a slower model. It can handle a lot of different feature types, like binary, categorical and numerical.

Gradient Boosting

Anomaly detection where data is often highly unbalanced such as DNA sequences, credit card transactions or cyber security.It can be used to solve almost all objective function that we can write gradient out. Are more sensitive to overfitting if the data is noisy.Training generally takes longer because of the fact that trees are built sequentially.

Logistic Regression

Image Segmentation and Categorization, Geographic Image Processing, Handwriting recognition, Spam detection and so on. Simple algorithm that provides great efficiency, variance is low, can also used for feature extraction and can be updated easily with new data using stochastic gradient descent. Doesn't handle large number of categorical variables well, requires transformation of non-linear features and are not flexible enough to capture complex relationships.

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.

Project Design

- Programming language: Python 3.6
- Library: Pandas, Numpy, Scikit-learn, Matplotlib
- Algorithms (Support Vector Machine, Linear and Logistic Regression, Decision Trees, Neural Networks and so on)
- Workflow:
    - Understanding of the dataset: create basic statistics, perform cleaning and processing if needed.
    - Feature engineering: Create new features based on the original features.
    - 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.
    - Perform supervisioned model training on the given data to achieve the prediction
    - Run the algorithms predictions
    - Compare the algorithms predictions
