Captone Proposal Maniceet

March 11, 2019

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Maniceet Sahay March 11th, 2019

1.1.1 Domain Background

NCAA Basketball tournament also known as March Madness brings with it the challenge of predicting the bracket and although it is unlikely to correctly predict all the matches correctly, maybe with data and machine learning we can assign some probabilities to these matches. The motivation behind picking up this as a project is the Kaggle competition that takes place every year around NCAA, the capstone also serves as an incentive to deep dive in the competition itself.

1.1.2 Problem Statement

Every year 64 teams take part in the annual NCAA Basketball competition and everyone gives in their prediction on who will go on to win the competition and due to upsets no one can predict the perfect bracket, but leveraging machine learning and historical data we can try to get as close as possible in predicting who will win.

The goal is to engineer features such as 3-pointer conversion rates and see whether they can help us in determining which team will win and give an objective view to the problem rather than going by expertise itself and have a quantitative approach to predicting the winners rather than a qualitative which most of us have.

1.1.3 Datasets and Inputs

The dataset can be obtained here

The data contains all sorts of data with the seed of teams from 1985 to 2018 and compact results where every result from 1985 to 2018 and detailed resulst which highlight results from 2003 to 2018 including all sorts of statistics such as number of 2 pointers attempted which will help us in engineering new features.

1.1.4 Solution Statement

The model will be either a logistic regression or a weighted aggregation of multiple models which may or may not include SVM, Random Forest, XGboost. Depending on the logloss obtained on

cross validation scores we will decide whether a single model does better or an ensemble performs better.

The approach will be to use the various features in the data and engineer new features and leverage the power of machine learning to outperform the basic model which only uses the seeds and seed difference as features

1.1.5 Benchmark Model

The bechmark model will be a logistic regression model run on the dataset with the seeds of the teams as the only feature, a seed difference will also be added as a feature and we will compare this model to our developed model using logloss as a metric.

With this we are trying to prove that engineering features helps us better in the prediction of matches when compared to only the seeds which donot reflect the outcome necessarily.

1.1.6 Evaluation Metrics

The evaluation metric will be log loss
$Logloss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$
where

- n is the number of games played
- $\hat{y}_i$ is the predicted probability of team 1 beating team 2
- y_i is 1 if team 1 wins, 0 if team 2 wins
- $\log()$ is the natural (base e) logarithm

A smaller log loss is better. The use of the logarithm provides extreme punishments for being both confident and wrong. In the worst possible case, a prediction that something is true when it is actually false will add an infinite amount to your error score. In order to prevent this, predictions are bounded away from the extremes by a small value.

1.1.7 Project Design

The data files will be first thoroughly checked for outliers and missing values. For columns which have over 30% missing values we will drop those columns and the rest we will use multiple imputation techniques such as mean replacement, median replacement and predicting the values using other factors as features. The correlation and distribution of these variables will also be checked so to understand whether certain transformations such as log can yield better results.

The main data file will be NCAATourneyDetailedResults.csv upon which we will build by adding new features such as but not limited to * Assist Ratio * Balanced Scoring * Competitive Balance Ratio (CBR) * Correlated Gaussian Winning Percentage * Defensive Efficiency * Defensive Rebounding Percentage * Strength Of Schedule (SOS) * Turnover Ratio * Value of Ball Possession (VBP) * Wins Produced For a complete list of features you can refer to these metrics listed by nbastuffer

Some of these metrics we can get from the main dataset itself like Assist Ratio whereas metrics such as CBR will require leveraging the other datsets such as RegularSeasonsDetailedResults to fill in gaps for teams which have not always played in NCAA Tournaments.

Base model will logistic regression based purely on seed and seed difference as features, we would first run that get the logloss as our benchmark score to beat.

The original dataset will be divided into train, test and validation with 80% in train, 10% in validation and 10% in test. The data set will be split according to years and not as a stratified split. Example, years 2003-2016 will be in train, 2017 will be validation and 2018 will be test.

The validation set should guide us in the right direction whether our efforts in EDA and feature engineering are providing any improvements over the baseline model. The test set will be reserved for predictions only once when we are confident about our model's results.

This should eliminate any chances of overfitting and also will indicate whether we have any variance problems, thus necessitating the need to include regularization or more training data.

Finally the goal is that using analytical tools and machine learning should give us an objective view of whether data can provide answers to basketball match outcomes or is it all in the heat of the moment.