Starcraft Data Mining Project

Thomas Eliassen
Noah Blumenfeld
Bridger Fisher
Ian Sime

Overview of Starcraft

RTS: Real Time Strategy Game: “Like chess but being able to move all the pieces at the same time.”

Players battle each other by first gathering resources and then building up their army. Players must use strategy to build, create, and control their army.


A BEGINNERS
GUIDE TO
STARCRAFT 2

Dataset

Our data is solely focused on the build stage of the game that involves building bases, gathering resources, and building the army. Our model's goal is to be able to predict the player's final strategy based on what order the player chooses to build each different piece of their army.

Distribution: Before preprocessing

The hard part about this dataset is that there are a lot of zero values that do play a vital part in reading the data. This makes the distribution of the dataset a little strange with lots of zero values and then fairly normally distributed data.

Distribution: After preprocessing

The hard part about this dataset is that there are a lot of zero values that do play a vital part in reading the data. This makes the distribution of the dataset a little strange with lots of zero values and then fairly normally distributed data.

Preprocessing

- Remove outliers
- Remove null values

Split the dataset

- 80% Training Data
- 20% Testing Data
- Create our own data

Notes

The features are always randomly permuted at each split. Therefore, the best found split may vary, even with the same training data and max_features=n_features, if the improvement of the criterion is identical for several splits enumerated during the search of the best split. To obtain a deterministic behaviour during fitting, random_state has to be fixed.

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

Algorithms

- Decision Tree
- Random Forest
- K Nearest Neighbors
- Gradient Boosting
- Logistic Regression

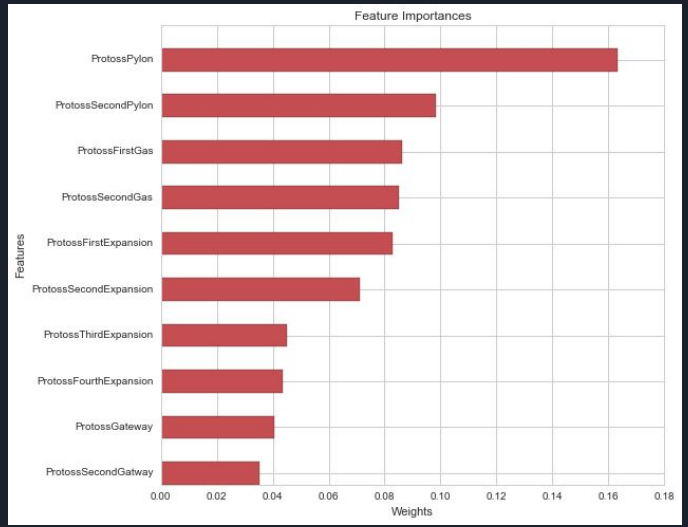
Model Results

Best Algorithm

Gradient Boosting

Best Feature: Bar Graph

This graph shows the importance of the features from the gradient boosting algorithm. The most important feature is ProtossGroundArmor1. ProtossGroundArmor1 describes starting the ground armor upgrade. Ground armor is used to protect ground units, essentially it makes it more difficult for them to be killed. The second most important feature is the ProtossDarkArchon feature. ProtossDarkAchron is a ground unit that supports other units. The purpose of the dark archon unit is to be a crowd control unit. It has the ability to take over the enemies' units, keep all units from moving, and the ability to drain energy from enemy units.



Best Feature: Pie Chart

This plotly pie chart shows the importance of features as well. In this chart the most import features is the ProtossThirdExpansion feature. ProtossThirdExpansion is referring to creating a third base and mineral field. This allows the player to increase the amount of minerals they are able to collect. Minerals are the currency in the game, you spend minerals to build and upgrade units. The second most important feature is the ProtossCannon feature. ProtossCannon is referring to the photon cannon structure. It is a defense structure that can target air and ground units. It is often used to defend “bases” that are on the map.

Categorical Scatterplot: Seaborn

The Seaborn Categorical Scatterplot shows the relationship of the occurrences of one attribute and its classifications at the midBuild, or halftime, of the game. The x-axis is the classification of strategy taken by the player at the midBuild, or halftime, of the game. The y-axis is the number of occurrences of that attribute. Then overall the scatter plot will show the relationship between the counts of each attribute and its classification. Therefore we can take each attribute and show the occurrences of each classification and look for patterns. The pattern we are looking for is if there are attributes that are most often categorized as a certain strategy or alternatively not very likely to be classified as a certain strategy.

Dynamic Histogram: Bokeh

The dynamic histograms we plotted show the distribution of the data for a given feature. Every graph has a large column for the 0 value, and a good distribution other than that. The reason that the 0s are still there after preprocessing the data is because they are extremely meaningful 0s. The 0s represent decisions that are made in gameplay. Choosing a different upgrade or unit creates a 0 in certain features. For example if the player chooses to upgrade ProtossGroundArmor1 the ProtossArmor1 feature would recieve a 0. If we were to replace these 0s with the mean for the data valuable information would be lost.

Resources

https://github.com/bgweber/StarCraftMining/blob/master/data/scmPvT_Protoss_Mid.csv

https://www.starcraft2.com/en-us/media#screenshots-13

https://youtu.be/2mwgriM2PM8

https://seaborn.pydata.org/tutorial.html

https://bokeh.pydata.org/en/latest/

Github

https://github.com/nblumenfeld/Data-Mining-Project