Sana Anwar, Myron Paes, Franchesca Vargas

Professor Monogioudis

DS 675-853

Video Presentation

Milestone 4 Report - Experiments on Catan Kaggle Dataset

The objective of this mini-project is to experiment with different machine learning concepts in this course for strategies to win the game Catan (via being the first player to obtain 10 points). The dataset *My Settlers of Catan Games* from Kaggle user Lumin provides a basis for all the game components – such as starting position, dice rolls, and resource trades – from 50 four-player games. From this dataset, there were over 20 dataset notebooks to start visualizing the patterns in the data between these Catan games. **[Sana]**

Kaggle user Matt Green performed *Logistic Regression and KNN* to analyze win/loss outcomes from early game statistics. The logistic regression showed that some of the features – including starting positions, resource production, robber cards gained, and cards lost within the game – have a strong linear relationship with the win/loss outcome, with an accuracy of around 82%. The KNN model achieved a slightly higher accuracy of around 84% when $K = 18$, indicating that these features might have a stronger relationship when nonlinear with the win/loss outcome. To improve the results of the logistic regression, finding the coefficients (β) of each feature will show which feature is a higher predictor of a win in Catan. The KNN model can be improved by conducting a grid or randomized search to ensure the optimal K is found. Based on our *Logistic Regression and KNN Colab notebook*, it was analyzed that the features with the most influence on the win/loss outcome include resource production (+1.887) and robber cards gained (+1.554). **[Sana]**

Kaggle user Abo Sol performed *Feature Engineering* to analyze the *combinations* of features that predict a player's final score. The most significant finding is a clear distinction between the original player (Kaggle user Lumin) and others, suggesting potential skill differences or dataset bias. Key insights derived from the features fall into several categories. In terms of Initial Position, access to Clay and Ore at the second settlement strongly correlates with higher scores. Regarding Game Style, high-scoring "me" players can still win despite poor initial tile probabilities by engaging in high-volume trade. The Game Duration insight shows that shorter games tend to penalize players other than Lumin, while longer games favor those skilled in resource management. For Resource Dynamics, players other than Lumin primarily obtain higher scores through production gains. Ultimately, Lumin's individual metrics are the strongest predictor, confirming the tracked player's higher average score (8.380 vs. 6.880). An improvement that can be done in the feature engineering is creating features for specific building pairs; for example, brick and wood can be paired for roads and settlements, while ore and wheat can be paired for cities and development cards. From this, a player's building capabilities can be another feature created based on these pairings. By doing *our own Feature Engineering* of these resource combinations, the synergy of ore and wheat (12) indicates that initially placing city/settlements on these resources to build more cities or development cards leads to a higher success rate. On the other hand, the synergy of having brick and wood (for settlements and roads) seems to be less important for success, as two winners of previous games had 0. **[Sana]**

Settlement placement is a foundational facet of Catan, and such decisions create “trees” of possible outcomes; as such, one can craft Random Forests stemming from settlement selection to attempt to accurately predict the subsequent final point totals of a user. Employing the *Decision Tree* technique (utilizing R), Abo Sol’s analysis primarily stipulates that sheer card quantity is positively correlated with victory point totals (*yval* within the attached notebook), insights which remain consistent with the findings of our aforementioned feature engineering notebook. Resource maximization, both through high-probability settlement placements and en masse trading practices, is critical to final point counts, ultimately augmenting a player’s chances of achieving ten victory points. **[Myron]**

Though the Logistic Regression conducted in this report already elucidates certain features which correlate to game outcome, the Random Forest foundation laid by Abo Sol enables this study to employ a Greedy Attribute Selection approach to deliberate on the *most* influential attributes within a game; we deployed a Greedy Attribute Selection Colab notebook with such an approach, ultimately selecting the eight most essential features of the Catan dataset. The results derived from our analysis indicate that, in chronology from most important to least, the features ‘7’, ‘5’, Total Loss, Total Available, Gained Cards from Robber, Total Gain, Trade Loss, and Second Player’s Second Settlement are all critical features of the Lumin’s Catan dataset, findings which align with all unearthed insights thus far. Features pertinent to card quantity (e.g Total Loss) comprise the bulk of the important features, reaffirming that a player’s card quantity is one of the most significant predictors of their end outcome. The features ‘7’ and ‘5’ may also be categorized as a quantity metric considering their high frequencies within the dataset; though the specific appearance of the two aforementioned numbers is a result of overfitting on Lumin’s Catan Dataset, this study can draw the overarching insight that maximizing card probabilities (known colloquially as “pips”) can also augment end victory points. **[Myron]**

There is a great overlap between our machine learning pipeline and the principles of ensemble learning, and in particular, we used Random Forests as the basis of our Greedy Attribute Selection algorithm. Random forests are an example of an old approach to ensembles. It is a technique that consists of a large number of decision trees, each being trained on a different subsample of the data, to give a more accurate and more stable prediction than any one tree. In our experiment, the random forest model was the tree averaging model, which was the best choice when it came to determining the features that always had an impact on Catan outcomes. The highest-ranked features (including 5 and 7 dice numbers, card profit/loss, and pips) were selected as they were found to be a good predictor by the ensemble model in its large number of internal decision courses. **[Franchesca]**

Ensemble learning was also integrated into our workflow by using Recursive Feature Elimination (RFE) on a Random Forest estimator. RFE systematically eliminates the most insignificant features and re-trains the ensemble model several times, similar to the iterative process of boosting algorithms. The repeated re-assessment of feature significance is not a boosting procedure in itself, although the repeated repetition of feature importance fits in with the boosting mentality, as the model improves with each re-evaluation. The result of this process was that the eight most influential attributes were isolated, and this shows how feature evaluation by an ensemble can help bring out underlying trends that linear models, such as the Logistic Regression, do not. **[Franchesca]**

Boosting, in principle, addresses the issue of fixing the error of the model in a series of rounds, and although we did not directly apply boosting to our notebook, the rationale of our analysis is nearly identical to that of boosting. Boosting models perform well in data where there are complicated interactions and non-linear relationships - just what occurs in Catan, where the flow of resources and probability of dice and strategic placement interact in subtle ways. RF has been used to reveal these trends, and a boosting model would probably highlight miscollected or hard/impossible to forecast games, like games where poor resource production has been compensated by good trading action. Therefore, our experiment preconditions how further enhancement of predictions may be achieved with the focus on edge cases. **[Franchesca]**

In particular, the findings of our ensemble-based feature importance are confirmed and supported by the results of Logistic Regression and KNN presented at the beginning of the project. Although the Logistic Regression found linear correlations, KNN found local structure, and the Random Forest ensemble ensured that the resource dynamics rule with the majority of the votes on hundreds of trees being correct. Such a cross-evaluation of several types of models is one of the characteristics of a successful ensemble methodology: to create a more stable picture of the data, other viewpoints are combined. **[Franchesca]**

Adopting the combination of Random Forests and RFE in our workflow allowed our project to directly leverage the strengths of ensemble learning while also introducing us to the core ideas behind boosting-based iterative refinement. Random Forests function by aggregating predictions from many individual decision trees, each trained on different subsets of the data, which enabled our analysis to capture complex interactions within the Catan dataset that single-tree or linear models could easily overlook. Because each tree brings a slightly different perspective and the ensemble aggregates its insights, this approach provides a more stable and reliable assessment of which features truly matter for predicting game outcomes. **[Franchesca]**