Expansion Impact Analysis – Touch N Go Soccer Third Location

David Carbajal
Computer Science Department
California Polytechnic University
Pomona, United States
djcarbajal@cpp.edu

***Abstract*—This project investigates predictive modeling approaches, specifically used to forecast customer behavior across multiple business locations. Touch N Go Soccer, is a youth soccer training organization currently operating through two locations, with plans of expanding to a third. Opening a new location requires a deep understanding of customer behavior, and without data driven insights it becomes difficult for Touch N Go Soccer management to evaluate potential sites, and decide where expansion would be the most successful. Throughout the first half of this project, the purpose was not to implement a full solution to this problem at hand, but to do thorough research on existing modeling techniques used in these situations. Study the strengths and weaknesses of these techniques, and formulate a problem statement. The final outcome of this project is a detailed direction and approach to solving this problem at hand. These findings successfully set up the next half of the project, where the goal is to implement forecasting models that will estimate cannibalization, complementary growth, and expected customer turnout.**

I. Introduction

Touch N Go Soccer currently operates through two main facilities located in Southern California. More specifically, these facilities are located in Corona, CA and Tustin, CA. Plans to expand to a third location have been discussed, but in order to move forward with these ideas, it is a necessity that management understand how this expansion will influence customer attendance across all locations.

The key concerns consist of cannibalization, complementary growth, and total expected attendance. Cannibalization, will the third location pull customers away from the Tustin or Corona locations? Complementary growth, will the new location bring in new customers who currently do not attend because of traveling distance? Finally, total expected attendance, what will the overall demand look like at the new location?

In order to create accurate predictions, the analysis of historical attendance patterns, demographic variables, and geographical relationships become major factors. The goal is to build a well thought out foundation and plan, that will take into consideration the mentioned features when implementation begins.

II. Problem Statement

Numerous models being used today provide strong starting points for unraveling hidden information behind customer behavior. Although this is the case, no models designed for analyzing youth sports training facilities specifically exist.

Current research does not provide accurate methods of implementing behavior models with the integration of spatial competition, demographic variables, and behavioral attendance patterns.

Attractiveness metrics also prove to be difficult to quantify. Attractiveness of a location can come down to trainer quality, class offerings, schedule convenience, and so much more.

Historical data is also limited, directly making the data preprocessing stage much more difficult.

Lastly, competing youth soccer training facilities, recreational leagues, and private coaching businesses also prove to be a factor that influence customer behavior. Ignoring external competition would risk overestimating demand at a new location, therefore we must consider competing soccer businesses when working with any of the proposed models.

The core problem at hand is the following: How can we build an interpretable, data driven model that integrates spatial competition metrics, demographic features, and past customer behavior? Specifically, how can we implement such a model to assess cannibalization, complementary growth, and customer attendance where Touch N Go Soccer considers opening a new location?

III. Literature review

1. Spatial interaction model: The Huff Model.

Spatial interaction models assist in estimating how a customer will choose between competing locations based on two key factors. Distance and attractiveness. The Huff Model is the most widely used spatial interaction model, and is one of the 3 routes being taken in this project.

Its strengths are that the model is very simple, easily interpretable, and widely validated in retail research. The Huff model also captures geographic competitiveness which is essential for predicting cannibalization.

The weaknesses of the Huff model are that attractiveness is subjective, and difficult to quantify. The model does not include demographic variables, and assumes customers decide only based on distance and attractiveness.

The Huff model will prove to be an essential way of estimating both internal cannibalization against existing Touch N Go soccer locations and competition from external youth soccer businesses.

Huff Model Formula

$$P_i = \frac{\frac{A_i}{D_i}}{\sum_j \frac{A_j}{D_j}}$$

Where:

- A_i: attractiveness of location i
- D_i: distance or travel friction
- P_i: probability a customer chooses location i

Fig. 1. Huff Model Formula

2. Regression Forecasting Models

Regression models relate attendance outcomes to specific variables that include but are not limited to distance, household income, youth population, school proximity, seasonal patterns, and much more.

The strengths of the regression models are that they are easily interpretable, they identify which variables matter the most, and are useful for estimating total expected sales and attendance at a new business location.

The weaknesses of the regression model include its lack of ability to find non-linear relationships. The model requires extremely clean and stable historical data, and does not capture spatial competition.

Regression will allow us to understand what factors play the largest roles in predicting attendance at both Tustin and Corona. With this information, forecasting new demand will be possible through the comparison of the new locations demographics.

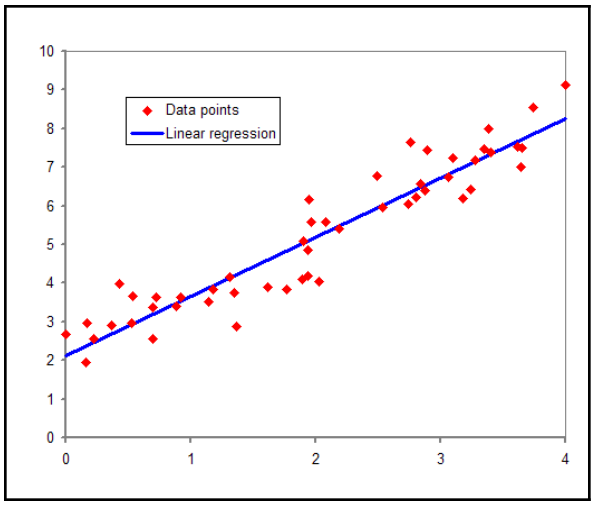

Fig. 2. Regression Model Example

3. Machine learning Models: Random Forest and XGBoost

The use of the Random Forest and XGBoost models will give us the ability to identify non linear relationships between location demographics and customer behavior the regression model may have missed.

The strengths of these models are that they handle complex, nonlinear patterns. These models will also provide its own variable importance rankings, giving another point of view at how important specific demographic variables are. Historically, these models are more accurate than linear models.

Weaknesses of the Random Forest and XGBoost models are that it requires large amounts of cleaned and preprocessed data, and can easily overfit when training. These models are also less interpretable compared to the regression and Huff models mentioned before.

Machine learning models have the ability to identify hidden patterns and relationships amongst data. By the identification of these hidden patterns, machine learning models can strongly forecast attendance at a new Touch N Go soccer location.

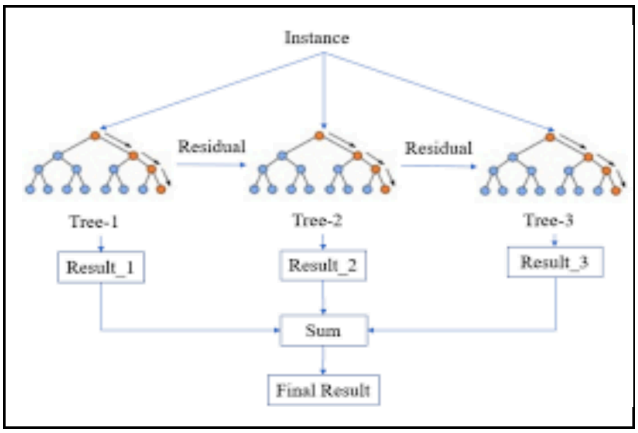

Fig. 2. Random Forest Visualization

IV. Implementation Strategy

How will these models be implemented into a process of determining competitive effects and possible location successfulness?

To determine cannibalization, both the Huff, regression, and Machine learning models will be utilized. Internal competition created by a new location will be determined by comparing pre-opening attendance metrics to model-predicted post-opening behavior.

The Huff model will be used as a baseline estimator for switching behavior amongst the customers. Using customer locations and distance to each Touch N Go location, the Huff Model will output a probability distribution of where each customer will most likely attend. Instances where customers whose predicted probabilities shift to the new location, represent cannibalization. Customers who predict to stay at their previous location, will represent retained demand.

Regression modeling will be applied to determine how demographic and situational variables contribute to customer attendance outcomes. By fitting regression models on historical attendance and demographic data, we will have the ability to identify which features drive customer attendance. These newly created coefficients will be used to estimate expected attendance of a new Touch N Go Soccer location.

The machine learning models will be implemented similarly to the regression model in the fact that predictive simulations will be run to identify major attendance drivers. The output will estimate expected customer attendance at a new facility location, using past data from the two existing locations.

V. Expansion Impact Evaluation

In order to evaluate how successful a new location would be, comparative performance frameworks would be developed that will integrate outputs from the Huff Model, regression forecasting, and machine learning predictions. The official goal is to determine whether or not the total demand increases after expansion, or whether or not the new location will merely redistribute existing customers. We will determine this with the use of estimated customer attendance predictions for the regression and machine learning models, and subtract this number by cannibalized attendance. The end result will deliver the overall successfulness of the new location.

VI. New Location Effectiveness

In order to evaluate how successful a new location would be, comparative performance frameworks would be developed that will integrate outputs from the Huff Model, regression forecasting, and machine learning predictions. The official goal is to determine whether or not the total demand increases after expansion, or whether or not the new location will merely redistribute existing customers. We will determine this with the use of estimated customer attendance predictions for the regression and machine learning models, and subtract this number by cannibalized attendance. The end result will deliver the overall successfulness of the new location.

VII. Project Timeline

The remaining work is the entire implementation process. In order, I will need to collect and preprocess all the data necessary, build the spatial baseline model, build regression forecasting models, build machine learning models, conduct scenario analysis, implement postprocessing techniques to visualize and explain the results, and prepare a final report and presentation.

VIII. Conclusion

This semester built the foundations for forecasting the impact of opening a new Touch N Go Soccer location. Through examining spatial models, regression models, and machine learning models, key insights into customer behavior will be found and explained. The next semester will focus entirely on implementation of these techniques, and results that will allow management to make comfortable expansion decisions. With this strong foundation set, the project is well positioned to have a successful second phase.

IX. References

1. [1] D. L. Huff, “A probabilistic analysis of shopping center trade areas,” *Land Economics*, vol. 39, no. 1, pp. 81–90, 1963.

[Online]. Available: https://www.jstor.org/stable/3144521?origin=crossref&seq=2

[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. OTexts, 2018. [Online]. Available: https://otexts.com/fpp2/index.html

[3] K. Kowsari, D. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. Barnes, "Improving retail store prediction through machine learning approaches," in *Proc. 25th ACM Int. Conf. Information and Knowledge Management*, 2016. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[4] G. Zhao, Z. Li, and Y. Liu, "Store site selection based on machine learning approaches," *Expert Systems with Applications*, vol. 134, pp. 41–54, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419306852?via%3Dihub