A Machine Learning Approach to Predict Season Ticket Membership

Julian Daduica Anna Nandar Jay Mangat Yeji Sohn

2025-06-24

Table of contents

1 Executive Summary

This project addresses the need for identifying which single-ticket buyers are most likely to convert to season ticket membership. Season ticket holders provide stable, predictable revenue that forms the financial background of the team.

Our team developed a solution using a machine learning model that sorts data into certain groups based on multiple criteria, called a multiclass random forest classification model. Our model uses Ticketmaster data from the 2022/23 to 2024/25 seasons for its analysis, which involves extensive data cleaning, feature engineering (incorporating domain expertise), and the creation of customer preference profiles based on a multitude of factors such as transaction timing, frequency, seat choices, and attendance patterns.

The final data product classifies customers into four categories: converted, churned, stayed member, and stayed non-member. Recall is prioritized, recognizing that the cost of missing potential converters exceeds that of additional marketing outreach.

This solution enables targeted marketing campaigns and optimized resource allocation by transforming season ticket member conversion from an intuition-based to an evidence-based strategy. Our final data pipeline requires minimal intervention to use and can be updated annually with new customer data. This allows the pipeline and Canucks Sports and Entertainment's (CSE) goal to increase season ticket sales to be up to date with each new hockey season.

2 Introduction

In today's competitive sports and entertainment industry, identifying which single-ticket buyers are most likely to convert to season ticket membership represents a critical business challenge. This identification is vital because season ticket holders provide stable, predictable revenue that forms the financial backbone of sports franchises. Unlike the variable demand of single-game sales, season tickets generate guaranteed income before each season begins and represent customers with significantly higher lifetime value, often accompanied by additional spending on concessions, merchandise, and premium experiences.

Can we leverage previous ticket purchase behavior to predict which customers are most likely to purchase season tickets in future seasons? Traditional marketing approaches result in broad, untargeted campaigns with poor returns on investment. Our goal of accurately predicting each customer's conversion likelihood would enable CSE to implement personalized outreach efforts and focus resources on prospects with genuine conversion potential. This targeted approach would enhance both financial efficiency and customer experience.

This analysis utilizes transaction data from Ticketmaster, comprised of detailed customer purchasing behavior across seasons 2022/23 to 2024/25. The dataset includes transaction timing, frequency, seat selection preferences, and event types alongside other features. Through initial data exploration and consultation with business partners, the original clustering-based problem statement evolved into a more targeted supervised learning approach.

The objective is achieved through a supervised multi-class classification approach using the Random Forest method. This choice aligns well with business needs for several reasons (Aljifri 2024). First, the method produces probability estimates for each customer, enabling ranking capabilities and flexible threshold adjustment. Second, the model's robustness to feature interactions and mixed data types suits the complex, multi-dimensional nature of customer behavioral data.

The evaluation strategy emphasizes recall optimization, recognizing that targeted marketing communications are relatively inexpensive compared to the high value of acquiring new season ticket holders. This approach minimizes false negatives (missing potential converters) while accepting more false positives (targeting non-converters), aligning model performance with business realities where the cost of missed opportunities far exceeds the cost of additional outreach or the risk of customer unsubscribing from marketing communications (Iyer, Yao, and May 2024).

These objectives directly address the core business need for increased season ticket sales by providing a data-driven approach to prospect identification. The predictive model enables evidence-based customer targeting rather than intuition-based marketing, supporting more effective resource allocation and improved conversion rates from single-ticket buyers to committed season ticket members.

3 Data Science Methods

3.1 Data Cleaning

The data from CSE's Ticketmaster account had several quality issues that needed resolution before analysis. A major step was filtering irrelevant or misleading entries. The dataset included not only sporting events but also concerts and entertainment shows. We also excluded preseason games, as they fell outside our focus on predicting season ticket members. We addressed missing data based on partner feedback and field context. For instance, most values in the home team column were null due to an export issue and were filled with "Canucks". We also cleaned the `userID` and `seat section` columns, as we intended to use them later.

A final key step involved cleaning the ticket type column. Though there were 91 unique values, most were rarely used or tied to one-off events and would not reappear. These were replaced with one of five main ticket types to reduce model complexity. The column also had incorrect entries where non-members used member-only ticket types, which we corrected.

3.2 Feature Engineering

Once the data was cleaned, we engineered features likely to signal whether a customer held a season ticket. We developed ours with input from our partners, mentor, and internal discussions to ensure they captured relevant behavior. Notable features included an ordinal variable ranking seating sections from 1 to 10 under the assumption that higher seat value correlates with membership likelihood. Combined with income data, this could signal purchase potential. Another feature tracked consecutive months attended, as full-season attendance often indicates potential members. We also restructured the response variable to reflect changes in membership status over time, making it more suitable for supervised learning.

Additionally, we introduced K-Nearest Neighbors (KNN)-based collaborative features as a way to add neighborhood-level behavior into each customer profile. Using cosine similarity over a selected set of behavioral features, we identified the five closest neighbors for each account. We then computed the average label distribution among these neighbors (e.g., percentage of neighbors who were season ticket members, stayed non-members, churned, or converted) and added these proportions as new features. This collaborative filtering-inspired approach enabled our model to better capture community or peer-like behavior patterns among similar customers.

The data cleaning and feature engineering stages resulted in unique customer profiles for each customer, which listed information about them from the original dataset and our new features. These profiles were converted to embeddings to be used in our unsupervised models and were used directly in our supervised models.

3.3 Multiclass Classification Using Random Forest

We chose Multiclass Random Forest Classification because it gives us a good balance between prediction and interpretability (Özen 2021). Our primary goal for this project was to categorize customers into one of four categories: `converted`, `churned`, `stayed_member`, and `stayed_non_member`.

The benefits of using this model are that it's capable of handling diverse data, as it can deal with both categorical and continuous features. It also allows multiclass classification support without extra modifications, which is useful in our case since we are predicting for four categories. Finally, random forest provides an importance score for each feature within the model, which increases transparency by allowing stakeholders to know which features the model finds most useful. This makes the model feel less like an artificial intelligence "black-box".

While this model has many positives, it has a slight drawback as well. Given its depth and complexity, it can take some time to train, and since we hope CSE will use this model for multiple additional seasons, this could add up. However, there are ways to work around this constraint by training the model on faster machines or employing cloud computing solutions.

Additionally, this model only needs to be trained once per season, so overall, this is a beneficial trade-off.

3.4 Other Machine Learning Approaches

Initially, we approached this as a binary classification problem to predict whether a customer would convert in 2025 (yes or no). We tried multiple models, including Logistic Regression, Random Forest, XGBoost, and Explainable Boosting Machine, but none of them gave us enough recall metrics — a recall of 0.70 to 0.90 would be acceptable, but none of these models exceeded 0.40 (Caruana 2020). Our dataset also had an imbalance issue, as only ~1% of customers converted to season ticket holders. Given these small groups, the models struggled to pick up on meaningful insights.

In an attempt to address the class imbalance, we tried oversampling and undersampling (Brownlee 2021). Oversampling duplicates the minority samples to get close to a 50/50 division of classes, while undersampling reduces the majority samples to get to 50/50. Both of these approaches failed to make a significant difference in model performance, hence we shifted our approach and treated the problem as a multiclass classification task instead. By including all customer behaviors like staying as a member, staying as a non-member, churn, or conversion, we gave the model more structure to learn from. This helped the model distinguish between different types of customers more effectively.

3.5 Evaluating Model Performance

After discussing with the Canucks team, we primarily evaluated our model's performance using Recall and the Confusion Matrix:

- **Recall**: Our main objective is to maximize the capture of true conversions, even if it means including some false positives. These false positives can still be valuable as potential leads for the marketing and sales teams.
- **Confusion Matrix**: This provides a breakdown of the model's performance across all classes, detailing true positives, true negatives, false positives, and false negatives.

Our model directly impacts several key stakeholders:

- **Marketing Team**: Model insights help them design targeted campaigns for potential season ticket members, ultimately increasing conversion rates.
- **Sales and Ticketing Team**: The model helps them prioritize high-potential leads, significantly improving the efficiency of their outreach and sales efforts.
- **Data and Analytics Team**: The model's interpretable outputs serve as a foundation for informed business decisions and can be leveraged for future model development.

3.6 Feature Importance

To enhance model interpretability, we examined the feature importance scores generated by our Random Forest classifier. These scores reflect how much each feature contributed to the model's predictive performance. Understanding which features the model relies on provides actionable insights for the business, enabling targeted improvements in outreach strategies and data collection.

Below, Figure 1 shows the most important features used by our Random Forest model to predict 2026 season ticket membership outcomes. Notably, the top contributors were the KNN-based collaborative filtering features (collab_prob_converted, collab_prob_stayed_member, collab_prob_stayed_non_member, and collab_prob_churned). These neighborhood-informed features dominated the model's decision-making, highlighting the strong influence of similar customers' behavior on individual predictions.

Beyond the collaborative features, recent engagement metrics such as events_attended_24 and Season_Ticket_24 also ranked highly, suggesting that the most recent season's activity provides key predictive value. Features like consecutive_months and avg_proximity had moderate importance, while seat quality (e.g., premium_section_ratio) and product type counts (e.g., Minis_24, Groups_23) had relatively lower importance. This distribution shows the effectiveness of incorporating both recent behavioral signals and peer-based context into the feature set.

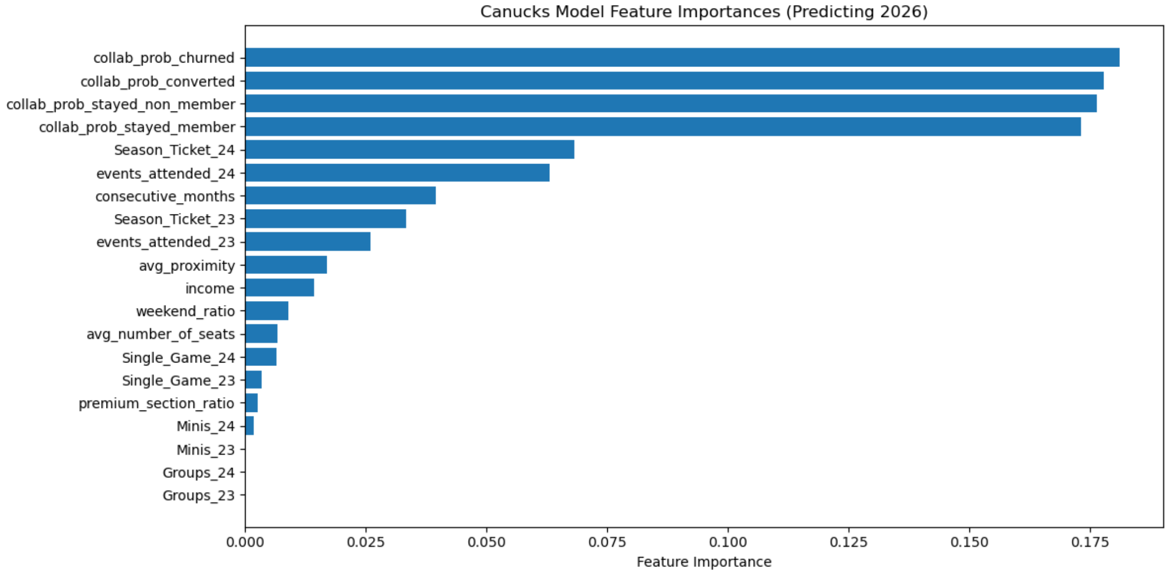

Figure 1: Canucks Top Model Feature Importances (2026 Prediction)

4 Results

4.1 Performance

The model demonstrates strong performance in identifying customers who convert to season ticket memberships (STMs), achieving a recall of 0.88 for the Converted category. This indicates that the model correctly identifies 88% of customers who actually convert to STMs, representing effective capture of true positive cases within this classification.

This high recall performance directly supports the primary business objective of maximizing conversion through accurate identification of prospective STMs. The model's ability to minimize false negatives is particularly valuable given the revenue implications of failing to target potential converters. The strategic emphasis on recall optimization aligns with the established business priority, where the opportunity cost of missing a potential STM significantly exceeds the resource cost of including non-converters in marketing campaigns.

While the precision for the Converted category is 0.43, indicating a substantial number of false positives, this metric trade-off is consistent with the project's strategic framework. The lower precision reflects the model's conservative approach to classification, which prioritizes comprehensive identification of potential converters over classification specificity. This approach ensures maximum market penetration while accepting the associated costs of broader targeting, ultimately supporting the organization's revenue optimization goals.

Table 1: Model Performance

Category	Precision	Recall	F1-Score	Support
Churned	0.45	0.83	0.59	52
Converted	0.43	0.88	0.58	152
Stayed Member	0.94	0.93	0.93	798
Stayed Non-member	1.00	0.98	0.99	13325

4.2 Overview

Our data product is a machine learning pipeline that takes in membership and transaction data from 2022 onward and outputs predictions for each account's membership status in the newest year. The model classifies each account into one of four categories: churned, converted, stayed member, or stayed non-member (see Table 2). The primary output of interest for the partner is the set of accounts predicted to convert, which are prioritized and converted in a separate output table (see Table 3). With a probability threshold 0.70, this allows the partner organization to focus efforts on the highest-potential leads for conversion.

Table 2: Example Customer Prediction Categories

Account	Predicted Label	Probability Churned	Probability Converted	Probability Stayed Member	Probability Stayed Non-Member
100110	Stayed Member	0.138	0.008	0.853	0.0
102167	Stayed Non Member	0.168	0.099	0.158	0.575
104752	Churned	0.516	0.0	0.483	0.0
14247275	Converted	0.0	0.715	0.013	0.272

Table 3: Example Accounts Predicted To Convert

Account	Predicted Label	Probability Churned	Probability Converted	Probability Stayed Member	Probability Stayed Non-Member
7878280	Converted	0.0	0.998	0.0	0.002
1234135	Converted	0.0	0.867	0.001	0.132
13719931	Converted	0.0	0.867	0.002	0.131
14203041	Converted	0.023	0.703	0.132	0.142

4.3 Intended Use by Partner

The partner can use the predicted list of high-conversion accounts to inform ticketing strategies, marketing campaigns, and personalized outreach. By narrowing the focus to accounts most likely to convert, they can more efficiently allocate resources.

4.4 Data Product

The primary advantages of our data product are the ability to support targeted prioritization, explainability, and scalability. By outputting accounts with a high predicted probability of conversion, the model enables more focused decision-making and increases efficiency in marketing and sales outreach. Additionally, because the model is based on a supervised learning approach, it provides feature importance outputs that make predictions interpretable. This level of transparency not only builds stakeholder trust but also offers actionable insights into the reasons behind conversion. The pipeline is also designed to be scalable, allowing it to be reused annually with minimal effort, simply by supplying updated data for the new season.

However, the product also has limitations. The model demonstrates moderate predictive accuracy, which increases the false positives; these are accounts flagged as likely to convert but that ultimately do not. While these accounts still share expected characteristics with actual converters, unfortunately, they do not, which decreases model performance.

4.5 Comparison Between Approaches

We selected a supervised learning approach (Random Forest Classifier) over unsupervised alternatives. Unlike unsupervised methods, our approach provides clear labels, performance metrics (e.g., recall and F1 score), and interpretability via feature importance. Unsupervised methods lack predictive targeting and require manual interpretation after running the pipeline, making them less actionable for our partner's use case. Moreover, unsupervised models are not well-suited for incorporating time-based or label-driven behaviours.

5 Conclusion

The business problem addressed was identifying which single-ticket buyers are most likely to convert to season ticket membership, enabling the partner organization to transition from inefficient broad marketing campaigns toward targeted, data-driven strategies. Our data product successfully transforms raw transactional data into actionable business intelligence by providing probability estimates for each customer's likelihood of conversion. This enables organizations to focus marketing resources on customers with the highest conversion potential while maintaining flexibility through adjustable probability thresholds. The solution meets the partner's key objective of identifying target customers through a scalable, low-effort annual approach. By optimizing for recall, the model prioritizes capturing as many potential season ticket holders as possible, accepting some false positives to avoid missing valuable conversions.

While the model achieved its primary objectives, several technical limitations emerged. The moderate predictive accuracy results in a large number of false positives, creating resource allocation inefficiencies and increased risk of customer unsubscribing due to over-targeting. The limited historical timeframe constrains model performance, and as additional years of data become available, the expanding feature set will increase model complexity and computational requirements.

From a business perspective, the model identifies who to target but provides limited insight into engagement strategies or optimal timing. Integration with the partner organization's existing systems falls outside the current scope. The solution relies on static annual predictions that do not capture dynamic changes throughout the season, providing no real-time insights. Additionally, the focus on historical purchasing patterns may overlook critical external factors such as team performance, economic conditions, or competitive offerings that significantly influence conversion decisions.

5.1 Refinements and Challenges

Future improvements to the data product could include threshold tuning based on cost-benefit analysis, allowing for more strategic selection of predicted converters by balancing precision and recall according to the potential return on investment per account. Additionally, incorporating temporal or sequential modelling techniques, such as recurrent neural networks or time-series-based transformers, could help capture more nuanced behavioural patterns that evolve over time. Another valuable enhancement would be the development of interactive dashboards, which would enable non-technical users to more easily explore and interpret model predictions.

However, these potential improvements come with trade-offs. Temporal models need more years' worth of data; they introduce increased complexity, require more intensive feature engineering, and often demand longer training times while reducing model interpretability. Similarly, building a user interface was deprioritized in this iteration to allow the team to focus on building a robust and reusable modelling pipeline within the available project timeframe.

5.2 Potential Improvements

There are some model improvements which we could not explore due to time and computational resource constraints:

- Threshold Tuning Based on Cost-Benefit Analysis: This would allow for more strategic selection of predicted converters by balancing precision and recall according to the potential return on investment per account.
- Deep Learning Models: Limited by computational resources. These models could potentially uncover more complex patterns within the customer data, but their training demands are substantial.
- Time-Series or Survival Analysis: Could provide a better understanding of when a customer is likely to convert, but this was not possible with the limited historical data.
- Development of Interactive Dashboard: This would enable non-technical users to more easily explore and interpret model predictions.

5.3 Further Steps

Moving forward, several recommendations emerge to enhance the solution's effectiveness. Implementing probability thresholds based on rigorous cost-benefit analysis will optimize the balance between precision and recall according to marketing budgets and conversion values. Incorporating more granular temporal modelling techniques will reveal when conversions are most likely to occur throughout the season. Expanding the feature set to include external factors such as team performance metrics and economic indicators will improve predictive accuracy. Establishing A/B testing frameworks will enable measurement of actual business impact compared to traditional approaches. Finally, implementing systematic tracking of marketing campaign effectiveness will create feedback loops for continuous model improvement and data collection refinement.

6 References

Aljifri, Ahmed. 2024. "Predicting Customer Churn in a Subscription-Based E-Commerce Platform Using Machine Learning Techniques." PhD thesis. https://du.diva-portal.org/smash/record.jsf?pid=diva2%3A1857189&dswid=-1532.

Brownlee, Jason. 2021. "Random Oversampling and Undersampling for Imbalanced Classification." https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/.

Caruana, Rich. 2020. "InterpretML: Explainable Boosting Machines (EBMs)." https://people.orie.cornell.edu/mru8/orie4741/lectures/Tutorial4MadeleineUdellClass_2020Dec08_RichCaruana_IntelligibleMLInterpretML_EBMs_75mins.pdf.

Iyer, Ganesh, Yunfei Yao, and Zemin Zhong May. 2024. "Precision-Recall Tradeoff in Competitive Targeting." https://faculty.haas.berkeley.edu/giyer/index_files/Precision%20recall.pdf.

Özen, Burak. 2021. "Multi-Class Classification on Imbalanced Data Using Random Forest Algorithm in Spark." https://burakozen.medium.com/multi-class-classification-on-imbalanced-data-using-random-forest-algorithm-in-spark-5b3d0af9b93f.