AI-based Event Location Services

Sakthignana Sundaram Somaskandan
Dublin City University
Dublin, Ireland
sakthignana.somaskandan2@mail.dcu.ie

A report submitted to Dublin City University, School of Computing for module CA699I: Topics of AI 2022-2023. I hereby certify that the work presented and the material contained herein is my own excerpt where explicitly stated references to other material are made.

ABSTRACT

The growing need for entertainment and better opportunities to socialise leaves event organisers with much pressure to think creatively for an engaging experience, especially in the post-covid era. The post-covid age where hybrid systems of work/study are becoming the norm. The existing event management systems are doing a good job of catering to event organisers' needs. However, this growing technology landscape desperately needs better tools and services. In this paper, I propose using leading AI techniques to recommend locations for new events and forecast supply chain requirements using Clustering and Random Forest techniques, respectively. My idea focuses on utility, performance, better end-to-end user experience and business sustainability. A viable business model is also discussed to market the product as a successful start-up.

1 INTRODUCTION

Several event management systems provide attendees and organisers with tools to enable them to make the most out of an event. The systems include features such as [1]:

1. Access to event information anytime and anywhere, even without Wi-Fi or data services
2. Real-time push notifications
3. Mobile brochure that can make instant updates even during an event
4. Allowing attendees to personalise their schedule and set a reminder for specific shows/sessions
5. Enabling attendees to actively participate in each session, answer live polling and share their thoughts on the session feed
6. Networking by browsing attendee profiles
7. Setting up a website for the event
8. Event promotion
9. Ticketing system
10. Generate name badges

However, event organisers are still left to research the best location for an event and the products and services that would be required for the event. With the advent of AI, these features are made possible by building a recommender system that narrows down a set of places for an event based on previously successfully held events. Additionally, better supply chain requirements can be forecasted based on historical data.

2 OBJECTIVES

The system will have two value propositions, and they are:

1. A recommender system that shortlists a handful of locations given an event type. The machine learning (ML) algorithm is an unsupervised learning technique called clustering.
2. Supply chain forecasting to predict demand for a set of products and services that are historically sold at a given type of event.

The system will grow slower when scaling the solution globally due to the need to procure our data, which requires time and money. However, if open quality data is available in certain countries, we could use it as it would also benefit the country's tourism department. Hence, there is an incentive for the government/local authority to collect and make such information public.

Another constraint of the system will be production monitoring of the decisions made by the algorithms. As with most machine learning models, there is a problem explaining how the system derived a solution due to its black-box nature. However, as the ML field is undergoing a massive shift and adoption in the industry, better tooling and monitoring systems will be developed in due course to aid in successfully deploying and utilising ML models in production.

3 FUNCTIONAL DESCRIPTION

The overall workflow of a machine learning model is illustrated in Figure 1. Sections 3.1 and 3.2 below detail the proposed approach for the two features of the system. The pros and cons of the approaches are also discussed.

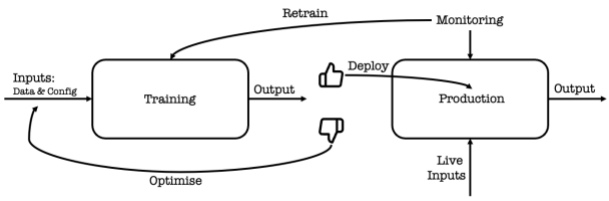

Figure 1: Machine Learning Workflow

3.1 Recommender System

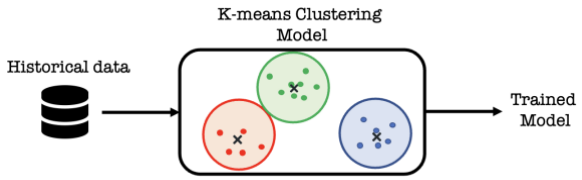

Figure 2: Recommender System using Clustering

The proposed recommender system uses the k-nearest neighbour approach to fetch and display several similar past events for a new event [2]. The dataset needed for the clustering algorithm would need to be compiled from the internet using various techniques like scraping and API usage to gather the required data. The data would need to be obtained from an event ticketing system, like Eventbrite [3]. However, the Activities dataset published by Fáilte Ireland can be a good starting point for the service [4]. The data-gathering phase is the most critical, challenging, and time-consuming part of the process, as the data quality dictates the system's success and accuracy. The following data points are required:

1. Date/Time of event
2. Event type/genre
3. Event location
4. Number of attendees
5. Total location capacity
6. Amenities around the location
7. Event rating

Once a substantial amount of data has been collected, it can be fed into the algorithm to partition the data into k clusters, with cluster labels being the event type. The feature set for the similarity measure will include the event type, location, amenities, rating, and time of year.

Upon successfully training the model, it will output a specific number of previously held events, which can then be ordered by the rating, the capacity or any filter made available to the user in the UI.

3.2 Supply Chain Forecasting

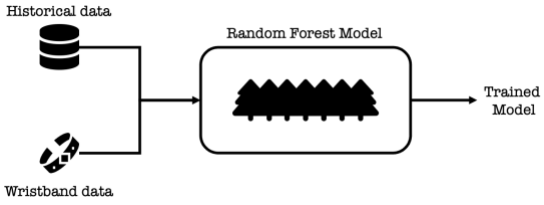

Figure 3: Supply Chain Forecasting using Random Forest

Supply chain forecasting, also known as demand prediction, is generally done using traditional statistical methods. These methods are prone to error due to their inability to capture patterns in the data appropriately. On the other hand, machine learning models will be able to learn inherent patterns and relationships in the data where the dataset could be a compilation of multiple different datasets, e.g., products' sales data could be combined with the weather data to draw a better picture of the scene. These patterns are derived directly from the historical demand dataset, such as seasonality and trend of the products.

The Random Forest ensemble algorithm can be used for forecasting, with the core idea being that there is wisdom in crowds. Each tree in a random forest is a decision tree capable of learning different patterns and relationships in the data compared to its sibling. Decision trees perform the classification or regression task by recursively asking simple true or false questions that split the data into the purest possible subgroups. In this ensembling method, several decision trees are trained, and the output from each tree is averaged as the random forest's output. The idea is that insight drawn from a large group of models is likely to be more accurate than any model's prediction alone [5].

The data required for the random forest algorithm would need to be collected from the hosted events. The products and services are sold inside the venue where the event is held, as it would be difficult to control and gather data from the surrounding area. Additionally, historical data can be gathered if available. Consequently, this part of the system would undergo slower growth. The data requirements are:

1. Attendance report for a show in an event.
2. Products and services purchased at an event.
3. Location data within the event premises – consented collection.

The above data requirements can be gathered by providing the attendees with a smart band. The smart band will allow users to record their experience and purchase products and services. It will have additional perks to entice the user to opt-in for it. The collected data would need to be cleaned in compliance with all required data protection regulations, such as GDPR, before storing it for analysis.

Over time, this data collection and analysis provides valuable insights into what products and services people gravitate to the most at a specific type of event, enabling future events to cater to the customer's expectations and demands accurately.

The advantages of using the random forest algorithm are:

- It works well with both categorical and numerical data.
- It allows for applying techniques like bagging and boosting to obtain a good bias-variance balance.
- It implicitly performs feature selection and generates uncorrelated decision trees by selecting a random subset of the available feature set for each decision tree, making it a great option for large datasets. It also allows for bootstrapping to sample the dataset for each tree randomly.

On the other hand, the disadvantages of using the random forest algorithm are:

- It can be hard to interpret the results obtained from a random forest.
- It is computationally expensive due to the requirement of building and training multiple trees as part of a single model.

The advantages of using the *k*-means clustering algorithm are:

- It is relatively simple to implement, especially when there is little knowledge about data distribution.
- It works well with large datasets.
- Generalises to clusters of different shapes and sizes.
- It is easily interpreted.
- It is computationally inexpensive.
- It is efficient.

On the other hand, the disadvantages of using the *k*-means clustering algorithm are:

- The *k* value must be manually selected – not optimal to preselect the number of clusters.
- It can only handle numerical data.
- It produces clusters with uniform sizes.
- It is sensitive to data transformations – normalising or standardising will impact the outcome.

The advantages of self-gathering the required data for the machine learning models are:

- It allows for the collection of quality data required specifically for the models' use case instead of using an open dataset which is only sometimes of the highest quality and intended for general purpose.
- It allows for the selection of features to collect.
- It is proprietary – not easy to replicate by other systems (companies)

The disadvantages of proprietary data gathering are:

- It is expensive.
- It is time-consuming.

4 EVALUATION PLAN

Several evaluation/performance metrics can be employed to quantify the trained model's quality. The proposed systems use different machine learning paradigms: unsupervised learning (recommender system) and supervised learning (supply chain forecasting). The *k*-means clustering is an unsupervised learning technique. Random forest is a supervised learning technique. The main difference between unsupervised and supervised learning techniques is the availability of labels. The supervised learning techniques usually require the ground truth labels to be available, whereas the unsupervised learning techniques don't require ground truth labels to draw insights/patterns from the data [6].

4.1 *k*-means Clustering

Accurately measuring the performance of a clustering algorithm is vital as it usually requires thorough inspection and validation. Therefore, a few different metrics would be used to determine the model performance [7][8]:

4.1.1 The Silhouette score measures the proximity of a point in a cluster to a neighbouring cluster. The score is in the range of $[-1,1]$, where 0 indicates that the sample is on or very close to the decision boundary; the closer the score is to 1, the further away the sample is to its neighbouring clusters' samples, and values less than 0 indicate that the sample is assigned to the wrong cluster. The Silhouette score can be used to determine the optimal value for *k* [9]. The Silhouette coefficient is defined as follows:

$$\frac{(b^i - a^i)}{\max(a^i, b^i)}$$

where

a^i is the average distance from all the data points in the same cluster

b^i is the average distance from all the data points in the closest cluster

4.1.2 The Calinski-Harabasz Index (C-H Index) – Variance Ratio Criterion – is used to evaluate the performance of a clustering algorithm without requiring the ground truth labels. The higher the index, the better the performance.

4.2 Random Forest

The random forest algorithm used for supply chain forecasting is a supervised regression model, which means the ground truth labels are required to train the model. It is essential to accurately measure the model's performance with metrics that communicate different aspects of the model: an error measure and a feature correlation measure [10].

4.2.1 Error measure – Root Mean Squared Error (RMSE) calculates the square root of the average of the squared error across all samples. This measurement reflects how spread apart the data points are from the regression fit. This representation of error has a significant advantage of being on the same scale as the target variable – easily interpretable compared to other error measures.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

4.2.2 Correlation measure – Adjusted R^2 (Coefficient of Determination) measures the relationship between two different variables in the model. This measurement gives the variation of the dependent variable that's directly related to the independent variable. The closer the value of R^2 is to 1, the stronger the correlation between the variables. The *Adjusted* part of the formula considers the more predictive features and gives them more weight than the less predictive features.

4.3 Production Evaluation

Besides the above evaluation metrics carried out during training, production evaluation plays a critical role in user engagement and experience, which drives value. Evaluating the results in production can be done in two ways:

1. Collect click-through rates.
2. Ask the user for feedback on the relevancy and effectiveness of the results.

5 DISCUSSION

The system would have the user flow as illustrated in Figure 4. The service's home page would have a few basic filters, i.e., event type and date/time of the event. The recommender system can then be queried using the entered filter values to fetch the nearest neighbours from the model. Subsequently, when the user clicks on a result, the details page displays all the relevant details regarding that event, e.g., the products and services sold at the location, attendee satisfaction etc. Alongside the event details, the forecasting tool would be available where the user (event organiser) could provide the number of anticipated attendees/sold tickets for their event and expect a quantity estimate of the required products and services at the event as an output.

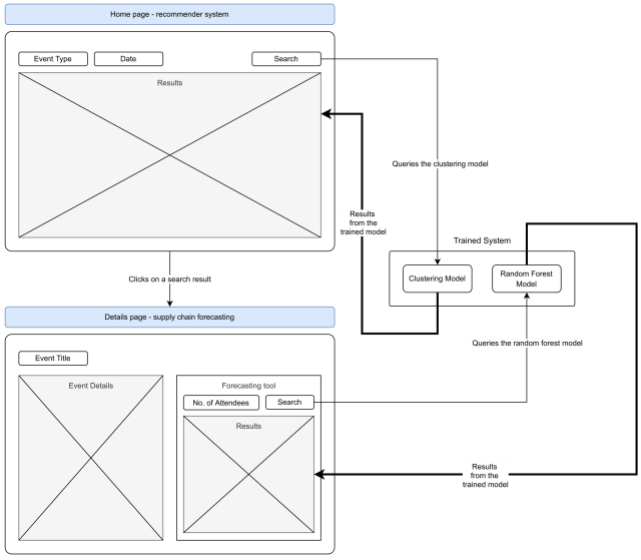

Figure 4: User flow diagram

This feature is not provided by any existing event management systems. Therefore, this service can be offered to existing event management systems as it complements their current services very well and provides an end-end experience for their customers, i.e., event organisers. The pricing would be a subscription model according to usage, as this service will be hosted in the cloud with API endpoints to query the required models.

This proposed service is better pitched to existing events management businesses as they have already established their business in the market. On the other hand, competing against existing players would require a custom implementation of all the services offered by the competition, which requires additional time and money upfront. Hence, taking a B2B approach for a start in the industry.

Once the service reaches maturity, other potential venues can be explored to expand the business and provide more value to existing systems as well as event organisers.

6 CONCLUSION

In conclusion, technology should be used to make people's lives easier, safer, and healthier and not solely to make big corporations more money. Even with an abundance of event management systems in the market, the event organiser's job is one of the most stressful jobs with a terrible work-life balance. There are plenty of opportunities in this space to build better systems to aid the event organiser's workflow. The ideas proposed in this paper are just a starting point for a great product with the primary goal of satisfying the information need of the workforce with affordance in mind to cater to a wide range of people.

REFERENCES

[1] "9 Event Management Software That Will Make You a Rockstar," *Whova*, 2022. https://whova.com/blog/free-event-planning-software-make-you-rockstar/ (accessed Dec. 08, 2022).

[2] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 971–976, 2010, doi: 10.1109/ICDM.2010.152.

[3] "API Reference | Eventbrite Platform," *Eventbrite*, 2022. https://www.eventbrite.com/platform/api (accessed Dec. 08, 2022).

[4] "Activities - Datasets - data.gov.ie," *Fáilte Ireland*, 2019. https://data.gov.ie/dataset/activities (accessed Dec. 08, 2022).

[5] N. Vandeput, *Data Science for Supply Chain Forecast*. Amazon Digital Services LLC - KDP Print US, 2018. [Online]. Available: https://books.google.ie/books?id=gbiRvgEACAAJ

[6] J. Delua, "Supervised vs. Unsupervised Learning: What's the Difference? | IBM," *IBM Analytics*, 2021. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (accessed Dec. 09, 2022).

[7] E. Zuccarelli, "Performance Metrics in ML - Part 3: Clustering | Towards Data Science," *Towards Data Science*, 2021. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6 (accessed Dec. 09, 2022).

[8] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, p. 052024, Jul. 2019, doi: 10.1088/1757-899X/569/5/052024.

[9] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 747–748, Oct. 2020, doi: 10.1109/DSAA49011.2020.00096.

[10] E. Zuccarelli, "Performance Metrics in ML - Part 2:

Regression | Towards Data Science," *Towards Data Science*, 2021. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-2-regression-c60608f3ef6a (accessed Dec. 09, 2022).