AI-based Event Management System

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

1 INTRODUCTION

2 OBJECTIVES

There are several event management systems that provide attendees and organizers with tools to enable them to make the most out of an event. The systems include features such as:

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 personalize 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 do their own research on finding the best location for an event and the products and services that would be required during an event. These features are made possible with the advent of AI, by building a recommender system that narrows down a set of places for an event based on previously held events. Additionally, better supply chain requirements can be forecasted based on the historical data.

How will it do it?

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 algorithm that will be used for this 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.

What constraints or limitations your system will have?

3 FUNCTIONAL DESCRIPTION

Give details of the techniques to be used.
Design a pipeline and provide the sketch here for each system
Also provide pseudocode for the 2 algorithms proposed

The overall workflow of a machine learning model is illustrated in Figure 1.

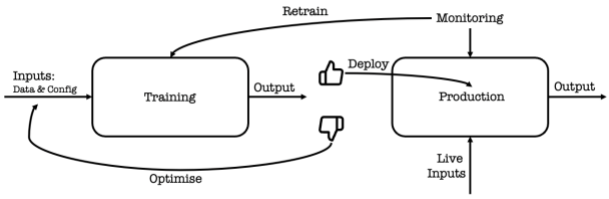

Figure 1: Machine Learning Workflow

3.1 Recommender System

The proposed recommender system uses the k-nearest neighbour approach to fetch and display the k nearest events for a new event. The similarity measure would make use of the event type/genre and time of year in the dataset. The dataset needed for the clustering algorithm would need to be compiled from the internet by using a variety of techniques like scraping and API usage to gather the required data. The data would need to be obtained from an event ticketing system, i.e., Eventbrite. However, the Activities dataset published by Fáilte Ireland can be used as a good starting point for the service. The data gathering phase is the most important, challenging, and time-consuming part of the process, as the quality of the data essentially dictates the success and the accuracy of the system. The following data points would be 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 type, location, rating.

Upon successfully training the model, it can be used to output a specific number of previously held similar events which can then be ordered by the rating, the capacity, or the date/time of the event according to the user preference.

3.2 Supply Chain Forecasting

Supply chain forecasting, also known as demand prediction, are 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, i.e., 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 task of classification or regression by recursively asking simple true or false questions that split the data into the purest possible subgroups. In this method of ensembling, several decision trees are trained and the output from each individual tree is averaged as the random forest's output. The idea being that insights drawn from a large group of models is likely to be more accurate than the prediction from any one model alone.

The data required for the random forest algorithm would need to be collected from the hosted events. 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. Real-time 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 the user to record their experience and purchase products and services with it. It will have additional perks associated with it 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 for analysis.

This type of data collections and analysis over a period of time provides valuable insights into what products and services people gravitated to the most at a specific type of event, enabling future events to cater for the customers' expectations and demands accurately.

Include an analysis of the pros and cons of your design.
Pros and cons of random forest – take from video

The advantages of using the random forest algorithm are:

- It works well with both categorical and numerical data.
- It allows for the application of techniques like bagging and boosting to obtain a good bias-variance balance.
- It implicitly performs feature selection and generate 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 randomly sample the dataset for each tree.

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.

Pros and cons of clustering… ?

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

- It is relatively simple to implement.
- 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 size.
- It is sensitive to data transformations – normalising or standardising will have an impact on the outcome.

Pros and cons of data gathering phase… ?

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

- It allows for the collection of quality data required for the models as opposed to using an open dataset which is not always of highest quality and intended for general purpose.
- It allows for the selection of features to collect.
- It is proprietary.

The disadvantages of proprietary data gathering are:

- It is expensive.
- It is time consuming.

4 EVALUATION PLAN

Explain how the system would be assessed and evaluated.

There are a variety of evaluation/performance metrics that can be employed to quantity the quality of the trained models. The two 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. The 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.

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:

4.1.1 The Silhouette score is used to measure 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' sample; 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*. 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 important to measure the performance of the model accurately with two metrics that communicate different aspects of the model: an error measure, and a feature correlation measure.

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 to the regression fit. This representation of error has a major advantage of being in 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 takes into account the more predictive features and gives it more weighting compared to the less predictive features.

4.3 Production Evaluation

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

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

5 DISCUSSION

How the system could form the basis of a successful business – or – a discussion of the architectural and system design aspects.

System design diagram

The system would have the user flow as illustrated in Figure 2. The home page of the service would have a few basic filters, i.e., event type and date/time of 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 would display all the relevant details regarding that event, e.g., the products and services sold at the location. Alongside the event details, the forecasting tool would be displayed where the user would be able to provide the number of attendees for their event and expect an estimate of the required products and services at the event as an output.

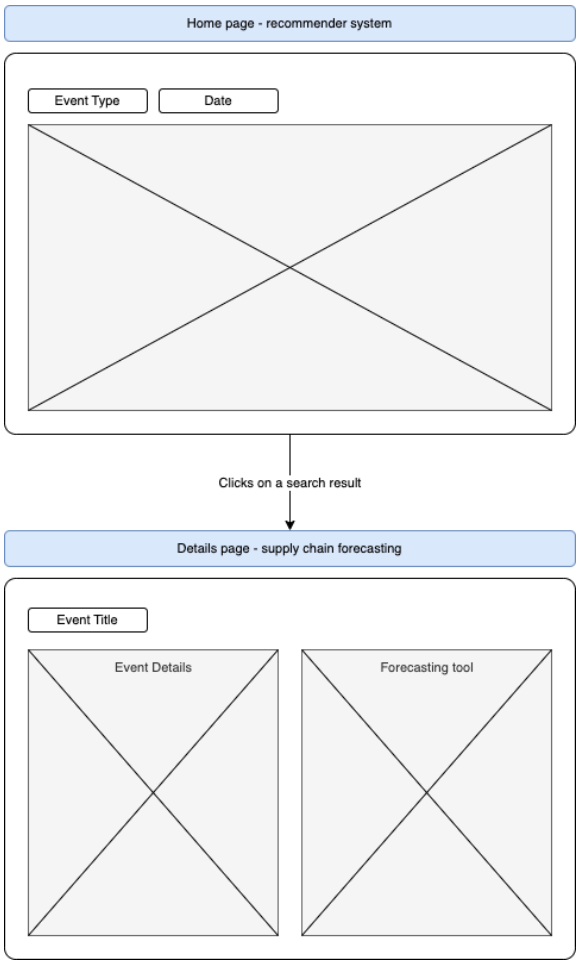

Figure 2: User flow diagram

This feature is not provided by any existing event management systems. Therefore, this service can be offered to event management systems as it complements their existing services 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 businesses in the event management space as they have already established their business. On the other hand, competing against existing big players would require a custom implementation of all the services provided by the competition, which in turn requires additional time and money upfront. Hence, taking a B2B approach for a start in the industry.

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

6 CONCLUSION

REFERENCES