Degree Project in Technology

First Cycle 15 Credits

Predicting Attendance at Events Within the Life-Science Sector Using Machine Learning

ALMA BAYNE, JOHANNA INNALA

Predicting Attendance at Events Within the Life-Science Sector Using Machine Learning

A. Bayne and J. Innala, *KTH Royal Institute of Technology*

***Abstract*—This thesis aims to apply different machine learning models on the complex problem of event curation to maximise attendance and customer engagement. Utilising historical attendance data and various demographic parameters from two different datasets provided by the life science industry organisation SwedenBIO, the study aims to provide event planners with reliable parameters that may affect a persons attendance. Testing the machine learning algorithms XGBoost, Random Forest and logistic regression to determine which of these gives the most accurate predictions on the two datasets as well as investigating which features each model deems as important for the decision making process. The findings demonstrate different strengths and weaknesses withing the models, but XGBoost gave the most accurate prediction for each dataset, exceeding the greatest from baseline comparison of both datasets. Features that impacted attendance the most for the Event-data were generally work titles as well as whether the event fell under the 'Knowledge' category. However for the Expo-data the models deemed company type as more important for deciding the type of participation with 'Investment', 'Biotech' and 'Service and Collaboration' in the top three, respectively. The findings of this thesis are able to assist SwedenBIO's event organisation in a way that boosts their event attendance and customer engagement.**

***Sammanfattning*—Den här rapporten ämnar att tillämpa olika maskininlärningsmodeller på det komplicerade problemet som är eventorganisering, för att kunna maximera deltagande och öka kundengagemang. Genom att använda historisk närvarodata och olika demografiska parametrar från två olika dataset som tillhandahållits av life science branschorganisationen SwedenBIO, avser studien att ge eventplanerare tillförlitliga parametrar som kan påverka en persons närvaro. Rapporten ämnar även att testa maskininlärningsalgoritmerna XGBoost, Random Forest och Logistisk Regression för att avgöra vilken av dessa som ger en mest exakt prognos på de två dataseten samt undersöka vilka egenskaper varje modell anser vara viktiga för beslutsprocessen. Resultaten visar på olika styrkor och svagheter hos modellerna, men XGBoost gav den mest exakta förutsägelsen för varje dataset som övertäffade baslinjen i båda dataseten. De parametrar som hade störst påverkan på deltagandet för Event-datan var yrkestitlar samt om eventet klassades som ett 'Kunskaps'-event. För Expo-datan påvisades istället att företagstyp var viktigare för att bestämma deltagandetyp med 'Investering', 'Bioteknik' samt 'Service och tjänsteorganisationer' som de tre främsta, i den ordningen. Resultaten från denna rapport kan hjälpa SwedenBIO:s eventorganisation på ett sätt som ökar deras evenemangsdeltagande och kundengagemang.**

***Index Terms*—XGBoost, Random Forest, Logistic Regression, Customer Engagement, SwedenBIO, Machine Learning, Attendance prediction**

We would like to express our gratitude to everyone at SwedenBIO for welcoming us over the course of writing this thesis and a special thanks to Maja Neiman, PhD, at SwedenBIO for her reliable support and consistent guidance throughout the writing process.

I. Introduction

A. Background

In practice, the way event planners curate invitation lists is often by using attendance data and relying on the intuitive judgement of planners regarding the perceived relevance of events for companies and their personnel. However, this approach may result in missed invitations for interested professionals, unnecessary invitations to unlikely attendees, and difficulty discerning the factors influencing attendance rates across different groups and events. Not inviting interested companies to events could deprive them of valuable networking opportunities, while aimless invitations may overwhelm potential attendees and diminish the perceived value of the event. Understanding the parameters that influence company and professional attendance at different types of events would enable planners to more efficiently tailor future events to specific groups or segments. A more efficient way of curating and marketing events to fit the target audience can boost the customer engagement. An emerging solution to address this uncertainty and avoid any guessing is the integration of data-driven insights into the invitation process. Incorporating past attendance data and predictive analytics into the invitation process can provide event planners with a more systematic approach to curating guest lists, ensuring that invitations are targeted and relevant, and maximising the potential for meaningful attendee engagement.

SwedenBIO is a Swedish industry organisation advocating for the interests of life-science companies, primarily in the biotechnology and medical technology sectors. As of March 2024 SwedenBIO has approximately 330 member companies and organises a variety of events, predominantly for their members but also for institutions and companies related to the life-science sector.[1] Their events give companies and representatives a chance to cultivate stronger business networks within their field and fostering greater exchange amongst industry peers - interactions that not only facilitate collaboration and knowledge sharing but also contribute to the advancement of the life-science industry. These interactions are made possible by their members and other professionals affiliated with the industry engage with their services. It is therefore imperative that companies receive invitations to relevant events tailored to their interests and needs, ensuring a high likelihood of attending and meaningful engagement from participants.

B. Study Objective and Research Purpose

The objective of this thesis is to apply machine learning models to the problem of curating invitation lists for different categories of events and analysing their effectiveness. By utilising invitations and attendance records from earlier events by SwedenBIO as well as data regarding invitees and respective company this study aims to find the most effective model for predicting attendance for different company representatives.

Additionally, it seeks to analyse important parameters that have a role in deciding whether an individual will, or will not, attend an event. This can help to ensure that SwedenBIO is using the correct marketing tactics and reaching the correct audiences for each event to maximise the engagement from their target audience.

C. Research Questions

- How can a machine learning algorithm predict if an invitee will attend an event within the life-science sector?
  - What machine learning algorithm is the most effective at performing such a prediction?
  - Is it possible to define important parameters, using machine learning, that are helpful for maximising customer engagement? If so, what parameters are relevant to predict whether or not a person will attend?

D. Societal and Ethical Aspects

The research conducted in this thesis holds relevance for SwedenBIO and its member companies, which are essential to the life-science industry in Sweden. By enhancing our understanding of how key figures in the industry come to connect and interact at SwedenBIO's events, the conclusion of this thesis can assist in making more of the right connections that contribute to the development of the Swedish life-science industry. The industry which is heavily dependent on collaborations to drive innovation, meaning that networking and relationship-building at events significantly impact the development of future medical innovations. As a result, this thesis additionally indirectly benefits society at a larger scale by optimising the networking and collaborative potential at these events within the life-science industry. Through its scientific insights and practical applications, this thesis showcases the capability of machine learning to further societal progress, especially in the vital area of health sciences, thereby improving public health and well-being.

The data used in this thesis has been collected and stored by SwedenBIO in accordance with GDPR, all individuals have previously accepted the terms and conditions of saving and processing their data for the purposes of providing them with relevant information and customised offers.[2] All personal data has been kept and processed in online spaces owned by SwedenBIO and not shared with anyone outside of the company.

E. Limitations and Scope

To ensure the scope of this thesis does not exceed the expected time-frame the analysis is only focused on the selected models deemed appropriate and reasonable for this project. Though the chosen models have previously proved to be useful in similar contexts it is important to acknowledge they may not capture all nuances of event attendance behaviour.

II. Theory

A. Logistic Regression

Logistic regression is a commonly used model in machine learning that falls under the category of supervised learning algorithms. It's primarily used when modelling the relationship between a categorical dependent variable and one or more independent variables. It is a versatile model that can be applied to a wide range of problems such as event prediction. A logistic regression model predicts the probability that an observation belongs to a particular class or category. In binary logistic regression the outcome $\sigma(x)$ varies in the interval [0-1] and translates to a binary classification where the dependent variable has only two outcomes, the observation is categorised as either 1 or 0; true or false.[3] The probability of the observation falling into the positive category is modelled by the sigmoid function:

$$\sigma(X) = \frac{1}{1 + e^{-\theta^T X}} \tag{1}$$

The goal of logistic regression is to find the parameters θ that, as accurately as possible, predict what class new observations belong to:

$$\theta^T X = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n \tag{2}$$

Here X represents the the values of the dependent variables and θ represent the corresponding parameters mapping to that specific feature. The act of training a logistic regression model refers to calculating the parameters θ, using data. By analysing the parameters it's also possible to understand how chosen features affect the prediction. [4]

B. Random Forest

Random Forest is another common machine learning algorithm that uses multiple decision trees to reach a single result. Decision trees start with a simple question and proceeds with a series of questions to determine an answer. These questions create new nodes in a decision tree which all contribute to making the final decision. Decision trees are commonly used in supervised machine learning algorithms but can be prone to problems such as bias and overfitting.[5]

Random Forest is made up of a multitude of decision trees that together make a more accurate prediction, especially when the trees are uncorrelated with each other, and are less prone to overfitting. The flexibility and accuracy of the Random Forest model make it well-suited for various problems, such as predicting events. The algorithm generates a random subset of features which guarantees unique decision trees with low correlation. When using the Random Forest algorithm three main hyperparameters are chosen before training that includes node size, number of trees and number of features sampled. This is then used to solve for regression where the individual

trees are averaged. The output of the Random Forest algorithm is more complex than that of a single decision tree but provides a more reliable answer. [5]

C. Extreme Gradient Boosting (XGBoost)

XGBoost, an acronym for Extreme Gradient Boosting, is an open-source software library for machine learning that utilises distributed gradient boosting decision trees under supervision. Supervised machine learning is based on learning a model to find patterns in a dataset with labels and features to then be able to make a prediction of labels on a new dataset based on its features. [6]

Similar to Random Forest, XGBoost builds a model with multiple decision trees but they build and combine the trees in a different way. XGBoost uses gradient boosting to improve the accuracy of the predictive model by sequentially combining multiple weak models (decision trees) into a single strong model. In Random Forest trees are built parallel to each other meanwhile XGBoost builds its next tree after the previous one is finished and in doing so correcting the errors made by the previous trees. The correction is made using a gradient descent algorithm on the loss function that determines how to adjust the predictions to minimise errors. [6] The high accuracy and many built in functions make XGBoost suitable for an extensive range of problems and can easily provide information about feature importance.

D. Evaluation of Machine Learning Algorithms

A standard method for evaluating how well machine learning algorithms fare involves calculating the following metrics: accuracy, precision, recall, and F1 score. These metrics are based on the amounts of true positives, true negatives, false positives, and false negatives. True positives and negatives represent the data points that the model predicted correctly, whereas false positives and negatives are all of the incorrect predictions. Higher accuracy indicates better model performance. Precision and recall provide insights into the model's exactness and completeness, respectively. The F1 score however might be more insightful since it is comprised of a combination of precision and recall and is particularly useful when dealing with an uneven class distribution.[7] How these metrics are calculated is represented in the formulas below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Another common approach to evaluating the performance of a predictive machine learning model is by establishing a baseline model - a less complex algorithm serving as a benchmark against which models can be compared. The general idea is that a more refined and complex model should surpass the accuracy and performance of the baseline model. Commonly for classification algorithms a popular baseline model is a so-called 'Major class classifier'; built on choosing the most frequently occurring class for all data points to get a simple benchmark accuracy.[8]

E. Customer Engagement

Customer engagement is a newer concept that has vastly grown in the last decade. It focuses on the way a company can create a relationship with its customers to promote brand loyalty in a way that transcends the initial transaction. This is achieved through precise relevant marketing that encourages the customers' interaction to repeat and is often accomplished through a digital form of marketing.[9]

Customers can be categorised into internal (e.g. employees, owners) and external (e.g. businesses, consumers) groups, as well as current or potential. Customers are the object of study in CE whereas engagement is the studied behaviour. Engagement is in marketing often organised into three groups: cognition (e.g. concentration), emotion (e.g. enjoyment, interest), and conation (e.g. attendance and participation in brand-related activities), reflecting the complexity of interactions between customers and brands.[10] When studying event prediction, a conation type of engagement, mostly the external current and potential customers are relevant.

The increased demand for data-driven decision-making in today's digital era, where the ability to make informed decisions based on previous customer interactions, can significantly enhance marketing efforts. By using customers data in determining patterns in their interactions, companies can refine their outreach and engagement strategies to better match the interests and behaviours of their target audience. Which is why customer engagement is an important tool for companies today to utilise and can lead to more meaningful connections between the company and their customers as well as improved brand loyalty. [11]

III. Previous Studies

Utilising machine learning for predictive modelling and classifications is a widely adopted approach, extensively explored and documented in various studies before. However, applying this methodology to predicting attendance at events, especially within the life-science industry, is less common. Similar studies have been made although not with the same focus on corporate events or industry conferences, each contributing unique insights and findings.

For instance, the study by J.K. Nguyen et al. (2022) [12] and the research by O.J. Leong and R. Mafas (2020) [13] both explored the use of machine learning to predict event attendance. Nguyen et al. focused on live events, using historical attendance data from football tickets to compare models like logistic regression, Random Forest, and XGBoost. By utilising historical attendance data regarding football tickets, models were made to predict future individual attendance

based on features like weather, demographic and past attendance records. In their research they found that XGBoost outperformed logistic regression when it came to accuracy, recall and f-score, however the logistic regression model saw a higher precision. The same conclusions were drawn by O.J Leong and R. Mafas (2020) in their research regarding the models. J.K. Nguyen et al. (2022) stated that the Random Forest model performed slightly less effective than the XGBoost, but still outperformed logistic regression in recall and F-score. While they deemed the XGBoost model better at predicting attendance in this case, it is also highlighted in the study that choosing what model to use is not without making a trade-off between the different measurements. For example, the authors present that the logistic regression model saw the highest true negative rate at 80.9% out of all compared models, but also the highest false negative rate at 48.8%. In conclusion, this suggests that machine learning models can successfully be used to get a better and more accurate prediction of individual attendance, as well as a deepened understanding of the important features that contribute to higher attendance.

This finding is supported by the authors O.J Leong and R. Mafas (2020), presented in their study reviewing predictive algorithms for event attendance in event-based social networks by leveraging the abundance of user generated-content available on social platforms to generate and compare predictive models based on different machine learning algorithms. The paper explores the intricate dynamics influencing users' attendance at social events, with a focus on predictive algorithms and event recommendation systems within Event-Based Social Networks. The authors present significant predictors such as the presence of friends attending and event location preferences. Moreover, factors such as event timing and day of the week are presented as critical variables of event popularity. Machine learning techniques are implemented to enhance attendance prediction accuracy and understanding evolving user preferences. These methods offer promising new paths for event planners to optimise event planning, target marketing strategies, and foster participant engagement in social events. Through a multifaceted approach, the study presents a comprehensive framework for event attendance prediction, paving the way for more effective event planning and management strategies and models.

Similarly to these studies suggesting XGBoost as the most accurate for predicting attendance a thesis by J. Awad (2023) [14], different machine learning algorithms are also compared, aiming to predict event attendance and create a recommendation system for invitations lists. The study extensively explores various machine learning models for event attendance prediction and guest list recommendation. Beginning with a baseline model, performance was evaluated using metrics like accuracy, precision and recall. Different models were investigated, with logistic regression showing versatility. Similarly to the study by J.K Nguyen et al. (2022) XGBoost was presented by the author as the optimal choice, in this case specifically due to its adaptability and reliability. Overall, the study highlights the synergy that can be retrieved implementing machine learning models with event planning, providing data-driven insights for better understanding how data can aid in the curating of events.

In conclusion, previous research in predicting customer engagement and attendance comparing various machine learning models, has yielded results indicating that utilising models can optimise invitation lists and give planners more insights into important event factors. However, it's inconclusive what models are optimal and the choice might vary depending on the exact circumstances of type of event and available data. This offers a robust theoretical foundation and also room for further exploration in this niche. Multiple previous conclusions state that XGBoost generally performs the best, still there are benefits and weaknesses for all the models, and thus there lies an interest to further investigate their practical applications with a new perspective.

IV. Method

In order to investigate the aforementioned research questions an approach similar to those of the previous studies was implemented. The provided data from SwedenBIO was first studied in order to gain knowledge and insight into the possible ways to proceed with the algorithms. Appropriate features described by the datasets were used in the analysis of the chosen machine learning algorithms. The following step was to perform data cleaning and prepare the data into neat categories fit for machine learning. Additionally preparing the models to apply to the data, as well as test and evaluate them.

A. Data Description

There were two sets of data provided by SwedenBIO. A set of 20 smaller events hosted by SwedenBIO in the years 2019-2024. These range from report launches and seminars to conferences and dinners. The dataset contained a list of all invitees, whether they attended or not, as well as their company name and work title. Across all these events there were collectively 10 852 data points. This data will be referred to as Event-data.

The second set of data contained a two-day international conference/expo held by SwedenBIO annually from the years 2014-2023 (with an exception the year 2020 due to the ongoing pandemic). The dataset was much more extensive and consisted of personal information of each attending person including work title and company name. Across all years there were a total of 10 249 data points. This will be referred to as Expo-data.

Additionally SwedenBIO provided access to their member-registry where all member companies and their affiliates were listed with among other names, emails and work titles.

Key differences in the datasets included their collection methods. The Event-data consisted of all individuals invited to each event. Since those who did not respond to the invitation did not provide their information, this dataset contained many incomplete data points. The Expo-data however only contained those who registered to attend the event, therefore this dataset was much more sufficient for evaluating the models even though it lacked non-attendees.

Due to the differences in the datasets these could not be combined or considered in the same manner. The Event-data was able to showcase a prediction of attendance but due to

the limited amount of data, it proved to be insufficient in evaluating the machine learning algorithms. Thus the Expo-data was additionally used in order to provide a better analysis of the models although lacking the ability to predict attendance in the same manner as for the Event-data and instead used to predict type of attendance. Therefore these two datasets have been processed separately with their own objectives.

After analysing both datasets with the study objective in mind, it was concluded, in unison with SwedenBIO, that certain parameters would provide the most insight through machine learning. For the Event-data, these parameters are event type, work title, company type, and attendance status. For the Expo-data, the relevant parameters are work title, company type, and type of participation. These parameters can help understand how customers engage with the events.

B. Data Cleaning

All unnecessary columns were removed from both the Event- and Expo-data, such as practical ticket purchasing information and time stamps. Then all duplicates of the same person within the same event/year were removed.

Due to the undemanding nature of some of the events within the Event-data, all forms have not been filled out properly with missing details of importance for our models. To fill some of these gaps the email addresses have been cross referenced with the member registry to find work titles and company names. The member registry was additionally used to categorise each company into a company type. All data points were not able to be completed due to the incomplete form and/or missing from the member-registry. This resulted in the Event-data having 4264 data points, with columns for event, attendance status, work title and company types.

For the Expo-data the columns collected from each year into one document included work title, company type and badge category. Due to the more extensive form all applicants submitted there was less data cleaning needed since it was mostly complete. The badge category however contained what type of participation each person would have, examples given below, and some were marked as replaced, refunded etc. which were removed. In total the Expo-data contained 9271 data points after it had been cleaned.

C. Data Categorisation

To ensure the algorithms function effectively, it is not considered ideal to have a wide variety of feature types. For example, using thousands of different work titles could hinder accurate results during training, as the model may struggle to identify patterns within the datasets. Therefore all columns in both datasets were condensed into 2-6 categories of somewhat equal sizes.

The Event-data was categorised into five different types of events: Policy, Dinner, Mixer, Knowledge and Summit. Summit being the name of one of their largest annual events having so many data points it was granted a category of its own. The attendance status consists of ones and zeroes, one meaning attended and zero meaning did not attend. There were six types of occupational titles that can be combined: Leadership, Technical expertise, Investment, Communications & Relationships, Consultant and Other. The companies were divided into five groups: Biotech, Pharma, CRO/CMO, Medtech and Service & Collaboration.

The Expo-data had the same categories for the work titles and company types. Additionally type of participation at the expo was divided into five categories: Exhibitor, Investor & Sponsor, Presenter, Participant and Organisers & Media.

A closer look through a visual representation of the ratios between the different categories for each dataset can be found in the appendix.

D. Splitting

Splitting of the datasets was accomplished using the 'train_test_split' function from the 'sklearn_model_selection' module. This works by shuffling the dataset prior to splitting it into one training set and one test set. Specifically, the size parameter was set to 0.2, effectively allocating 20% of the dataset for testing and 80% of the dataset for training of the model.

E. Training of Models

The initial step of model training involved converting all data points within the datasets into vectors using one hot encoding. Specifically, for attributes such as company type and event, as well as participation type for Expo-data, this was done using the 'get_dummies' function from the pandas library, given that each data point was associated with only one company type, event type and participation type. For the role category, which could encompass multiple roles per data point, encoding was performed using 'MultiLabelBinarizer' from the 'sklearn.preprocessing' module.

Once encoded, the vectors representing events, company types, roles and type of participation were integrated and the status, already in a binary format, was assigned to the target variable 'y'. Subsequently, a model was constructed, hyper-parameters were selected and the models were trained utilising the respective built-in 'fit' functions. Regarding implementation, the XGBoost model was sourced from the XGBoost library, while the Random Forest and logistic regression models were both imported from the sklearn library.

F. Testing and Evaluation

With the trained models, accuracy was determined using the test data and built-in evaluation functions. Additionally, a comprehensive analysis was conducted utilising the 'classification_report', providing precision and recall metrics for all classes. Feature importance scores were extracted directly from the model using relevant functions provided by respective library. This provided a list of the features in order of most to least important factors in the classification process. The five most significant features from each model were collected. Five were chosen since compared to the total amount of features, 17 for the Event-data and 12 for the Expo-data, the top 5 features fall in the top half of the list. These features were analysed in

order to determine whether they gave meaningful insight into the types of people likely to attend which events and in what capacity, that can be useful in improving future marketing and event organisation.

V. Result

A. Event-data

For the Event-data dataset the accuracy presented in Table I was attained using the three different machine learning models. As seen the top accuracy was attained using XGBoost.

Model	Accuracy
XGBoost	**0.69**
Random Forest	0.65
Logistic Regression	0.64

TABLE I
Event-data accuracy

For each model a more detailed classification report was extracted and can be studied in Table II.

XGBoost			
	Precision	Recall	F1-score
Declined	0.71	0.85	0.77
Attended	0.63	0.43	0.51
Random Forest			
	Precision	Recall	F1-score
Declined	0.69	0.82	0.75
Attended	0.62	0.44	0.51
Logistic Regression			
	Precision	Recall	F1-score
Declined	0.73	0.60	0.66
Attended	0.50	0.65	0.57

TABLE II
Event-data Comparison of Models

XGBoost		
Ranking	Feature	Category
1	Knowledge	Event type
2	Technical Expertise	Role
3	**Leadership**	Role
4	**Investment**	Role
5	Summit	Event type
Random Forest		
Ranking	Feature	Category
1	**Leadership**	Role
2	Technical Expertise	Role
3	**Investment**	Role
4	Consultant	Role
5	Communications & Relationships	Role
Logistic Regression		
Ranking	Feature	Category
1	Consultant	Role
2	**Leadership**	Role
3	Knowledge	Event type
4	Other	Role
5	**Investment**	Role

TABLE III
Important Features for Event-data from Different Models

Using built-in functions the features were ranked in importance for each model. The analyses yielded slightly varying results for the different models, with a more detailed list of the top 5 features in Table III. Features in bold are the ones that are present in all three models top ranking.

Generally for the XGBoost model the roles were deemed of high importance, the company types were deemed of less importance to the model and event types were spread out both towards the top and bottom of the list.

In Random Forest the general ranking of importance was with roles at the top, event type in the middle and company type as deemed least important to the model.

In the logistic regression model the general ranking of importance was with roles at the top, event type in the middle and company type as deemed least important to the model.

B. Expo-data

For the Expo-data dataset the accuracy presented in Table IV was attained using the three different machine learning models. As seen the top accuracy was attained using XGBoost.

Model	Accuracy
XGBoost	**0.58**
Random Forest	0.57
Logistic Regression	0.27

TABLE IV
Expo-data accuracy

For each model a more detailed classification report was extracted and can be studied in Table V.

XGBoost			
	Precision	Recall	F1-score
Exhibitor	0.38	0.06	0.10
Investor & sponsor	0.61	0.41	0.49
Organisers & media	0.34	0.20	0.25
Participant	0.64	0.91	0.72
Presenter	0.0	0.0	0.0
Random Forest			
	Precision	Recall	F1-score
Exhibitor	0.0	0.0	0.0
Investor & sponsor	0.68	0.32	0.43
Organisers & media	0.0	0.0	0.0
Participant	0.64	0.56	0.71
Presenter	0.0	0.0	0.0
Logistic Regression			
	Precision	Recall	F1-score
Exhibitor	0.26	0.46	0.34
Investor & sponsor	0.49	0.41	0.48
Organisers & media	0.64	0.20	0.17
Participant	0.54	0.08	0.15
Presenter	0.21	0.47	0.29

TABLE V
Expo-data Comparison of Models

Using built-in functions the features were ranked in importance for each model. The analyses yielded generally similar results, with a more detailed list of the top 5 ranked features in Table Table VI.

XGBoost		
Ranking	Feature	Category
1	Invest	Company type
2	Service & Collaboration	Company type
3	Biotech	Company type
4	Pharma	Company type
5	CRO/CMO	Company type
Random Forest		
Ranking	Feature	Category
1	Invest	Company type
2	Biotech	Company type
3	Service & Collaboration	Company type
4	CRO/CMO	Company type
5	Pharma	Company type
Logistic Regression		
Ranking	Feature	Category
1	Invest	Company type
2	Biotech	Company type
3	Service & Collaboration	Company type
4	CRO/CMO	Company type
5	Pharma	Company type

TABLE VI
IMPORTANT FEATURES FOR EXPO-DATA FROM DIFFERENT MODELS

Generally the company types were deemed of high importance and the roles were deemed of less importance for all three models, with event type somewhere in the middle rankings.

VI. DISCUSSION

A. Baseline Comparison

An appropriate baseline model for evaluating the predictive performance of more complex models would be a model classifying all data points in the set as the most frequently occurring class. This method serves as a straightforward benchmark, requiring minimal effort and complexity while still providing a more accurate point of comparison than randomly guessing the predictions.

For the Event-data, this baseline model would entail classify all data points in the test as 'Declining', since the majority of individuals across the board tend to decline the event invitations. Using this model for classifications, an expected accuracy of approximately 60% would be achieved, reflecting the 60% decline-rate.

For the Expo-data, this baseline model would entail classify all data points in the test as belonging to the 'Participant', since the majority of individuals across the board are classed as 'Participants'. Using this model for classifications, an expected accuracy of approximately 52% would be achieved, reflecting the 52% participant-rate.

With this baseline model in mind, it's possible and valuable to evaluate the models accuracy and performance in comparison to both the baseline and each other.

B. Model Comparison

When comparing the performance achieved for the different models on the Event-data dataset, the results indicate a broadly similar level of accuracy across the models, with XGBoost slightly outperforming the other two and logistic regression exhibiting the lowers accuracy score as seen in Table I. Notably, all three models surpass the baseline accuracy, although by modest margins, demonstrating that the models outperform the less complex baseline model.

In terms of recall and precision metrics, as presented in Table II, XGBoost and Random Forest presented comparable performance, with XGBoost achieving marginally higher precision for both classes, slightly higher recall in classifying 'Declined' and slightly lower recall for classifying 'Attended'. This suggests that the tuning of the XGBoost model might be more conservative, aiming to reduce false positives on the expense of a greater amount of false negatives.

The logistic regression model displayed, Table II, significantly lower precision for 'Attended' and lower recall for 'Declined', albeit the precision for 'Declined' and the recall for 'Attended' were higher. This result indicates that the model likely tends to favour classifying as attended over declined compared to the other models. This results in a higher rate of false positive, meaning non-attendees were more likely in this model to be incorrectly classified as attending. This is also strengthened by the fact that the F1-score for 'Declined' is highest for XGBoost compared to the other two models, with logistic regression at the bottom indicating that the XGBoost model tends to classify the 'Declined' more accurately, whereas the logistic regression model favours the smaller category 'Attended' more.

For the Expo-data dataset the difference in model accuracy, presented in Table IV, is evident, with XGBoost and Random Forest at 58% ans 57% respectively, while the logistic regression model only achieved an accuracy of 27%.

Further examination of the detailed classification reports reveal that the logistic regression model generally distributed the predictions more evenly across all participation classes, contrary to Random Forest, as seen in Table V, which did not classify any data point as 'Exhibitor', 'Organiser & Media' or 'Presenter'. Similarly, XGBoost did not classify any data points as 'Presenter'. Generally, it seems as though, for both the Event-data and Expo-data, the logistic regression model is more evenly predicting the classes and putting larger emphasis on the classes that are statistically less likely. Contrary, XGBoost seems to strongly favour the most likely class, tending to neglect the less likely classes with a more conservative approach. The Random Forest model seems to fall somewhere in between the two other models approaches.

Comparing this achieved result with previous studies within the same field it's found that XGBoost was similarly the outperforming model in other studies. It's presented in the study by J.K Nguyen et al. (2022), the thesis by J. Awad (2023) as well as the study by O.J Leong et al (2020) that the XGBoost models were found to achieve the highest accuracy out of the compared models. The authors unanimously present this machine learning model to be the best suited for this type of problem. In comparison, J.K Nguyen et al. (2022) also implemented a solution with a Random Forest algorithm as well as a logistic regression model. The Random Forest model outperformed logistic regression, but still fell short of the accuracy achieved by the XGBoost model.

The results of these previous studies all respectively are

reflected in the results acieved by this project, with XGBoost outperforming in terms of accuracy, with Random Forest close behind and logistic regression lagging after the two.

C. Feature Importance

For both the Event-data and the Expo-data, there was a notable consistency in the feature importance ranking across all three models.

For the Event-data there were a total of 17 features. As seen in Table III, roles were shown to be significantly influential, with role 'Leadership' and role 'Investment' consistently appearing in the top five most important parameters across all models. Furthermore, event 'Knowledge' was highlighted as a top-five feature in two of the models, coming in at number one in the XGBoost model and third in logistic regression. The role of 'Consultant' was also prominent, ranking fourth in the Random Forest model and first in the logistic regression model. Lastly, role 'Technical Expertise' also appeared in two models top rankings, as second in both XGBoost and Random Forest.

However, some categories presented only model-specific importance, only ranking in the top five for one of the models. The event 'Summit' was only present as number five in the XGBoost model, role 'Other' was exclusively noted in the logistic regression model as the fourth feature and role 'Communications & Relationships' coming in as fifth for Random Forest.

For the Expo-data there were a total of 12 features. As seen in Table VI, the emphasis was predominantly on the company type as an important indicator of type of participant and role was deemed of less importance. This pattern was demonstrated in all three models, as only company types were noted as the top five highest ranked parameters, specifically with company type 'Invest' as number one in all three models.

Company type 'Biotech', 'Service & Collaboration', 'CRO/CMO' and 'Pharma' were all present in the top five across all three models, with varying ranking, but generally for all three models 'Service & Collaboration' and 'Biotech' were ranked as either two or three, and 'Pharma' and 'CRO/CMO' were ranked as either four or five.

D. Customer Engagement

To maximise customer engagement, specifically related to conation, i.e. attendance and participation in events, it's key to ensure that all relevant individuals are invited to events to maximise the attendance. With the obtained results it's likely not feasible to use the models for predicting what types of individuals would be interested in different types of events to curate invitation lists, as the accuracy is not very high. It's also clear that all three models have a higher F1-score for the class 'Declined', meaning they better predict who is not going to attend, then who is attending. From a customer engagement perspective this is not a positive as it's likely preferable to invite too many rather than too few.

In this case, studying the behaviours of external customers that currently partake in these events could help with future curation of events and understanding the important parameters can help reach new customer audiences.

The feature analysis, where the features of external customers that currently partake in these events, could bring insight into how to plan upcoming events to both keep the current customers but perhaps also reach new customer audiences. Utilising these results, in this case that roles play a larger part in attendance and company type plays a larger part in type of participation, it's possible to look closer at roles and attendance to understand what roles are more or less likely to attend what events.

VII. Conclusion

The models each had strengths and weaknesses depending on what's deemed most important. If accuracy is the goal, XGBoost is the most appropriate model as it scored highest in accuracy for both datasets. The notion that XGBoost was the model with highest accuracy was also consistent with some of the findings in the precious studies. If it's deemed more important to lessen the amount of false negatives, i.e. classifying attending people as non-attending, is higher prioritised than accuracy, Random Forest or logistic regression might be more fitting when looking at a dataset like the Event-data. This could be due to, for example, a higher cost in underestimating the amount of people at an event. The strikingly lower accuracy for the logistic regression model compared to the other two for the Expo-data suggests that this model is likely not a good fit for this type of dataset.

It's possible to some extent identify what parameters play a role in the attendance of individuals at events. From the Event-data dataset it's clear that role plays a larger part than company type in identifying if a person is attending or not, specifically if the person had role as 'Leadership', 'Technical Expertise' or 'Investment', as this patters was seen across the board in all three models respectively and these were at the top 5 of the 17 features consistently. As both the type of event being 'Knowledge' and a persons role being 'Consultant' was ranked highly in two models it's this implies to some extent that these parameters also likely play a larger part in deciding if a person attends or not. Event type 'Summit' and role 'Other' was only ranked as top five most important in on model and were both also one the lower half, therefor it's not possible with certainty to say anything about the importance of there parameters to the classification.

Looking at the results for the Expo-data dataset, it's clear that company type plays a larger part in deciding the participation type than the role an individual has since this is evident in all three models with company type at the top of the importance ranking. In all three models company type 'Invest' was deemed the most important, 'Biotech' and 'Service & Collaboration' at spot two or three, and 'CRO/CMO' and 'Pharma' at four or five, indicating that this is likely a representative ranking of the important parameters.

Given this analysis of the important features, its evident that some parameters are more influential in whether an individual will attend and event or not, and in that case the nature of their participation. Consequently, for a organisation like

SwedenBIO, focusing more on these key parameters when planning events and curating invitation lists is crucial as these are factors most likely to affect if an individual attends or not. This targeted approach can help with maximising customer engagement.

This analysis provides a solid foundation for SwedenBIO to incorporate a more data-driven approach when orchestrating events and with more data and fine-tuning of models even more effective and insightful models could be made, not only to maximise attendance rates, but also deepen the customer engagement, fostering a stronger connection and advancing the life-science sector. With the awareness that it's likely primarily an individuals role at a company that determines weather they attend or not, gives valuable insight into the decision process regarding attendance and suggests that this features is what should be primarily considered when planning events.

For now, the attained results from the models are likely not good enough for the models to be used for curating invitation lists, as an accuracy of 70% risks overlooking many potential attendees. For the models to be used in practice a very high accuracy would be needed as it's unwise to risk not inviting relevant individuals. However, the results for important features are able to lay the groundwork on what parameters to look at when planning events indicating that what role an individual has at an organisation or company likely plays a larger part than what type of event or company.

A. Future Studies

For future studies in the field it would be valuable to test the models on different types of data as well as similar, but more complete data, to compare the results. A merging of models to reap the benefits of the different approaches, or exploring other machine learning models, could also be valuable. Although it was found that XGBoost was the most fitting model for this type of dataset, the different models had differing strengths and weaknesses that might make them a better fit for other types of datasets.

Another large factor contributing to the results is the dataset that the models are trained and tested on. With more complete data and a larger dataset, with fairly equally sized categories, the results will likely improve as this aids in the training of the model's parameters. This would give a more accurate understanding of how he features impact the classification by the models. To improve the cleaning and categorisation of data it's also a possible alternative to make filling out some information mandatory when signing up for events and perhaps making predefined categories for participants to choose from instead of themselves writing in text. Data collection is complicated and often not without ethical conundrums, but investigating better ways to collect data regarding both attending and declining individuals it would be possible to more accurately predict who will and wont attend.

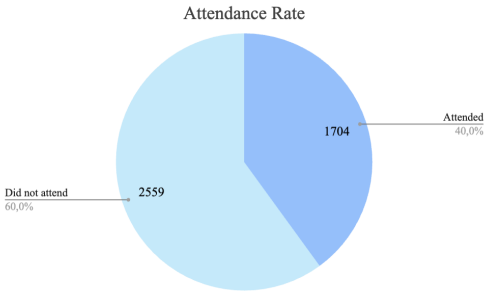
AUTHORS

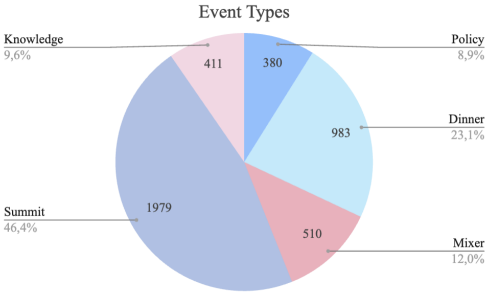
Alma Bayne and **Johanna Innala** are both students at the Royal Institute of Technology (KTH) in Stockholm, Sweden, currently studying Industrial Engineering and Management since 2021. They are the authors of this thesis and have collaborated closely on all parts, with both authors contributing equally to the final product.

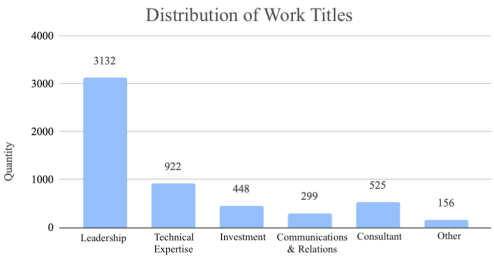
REFERENCES

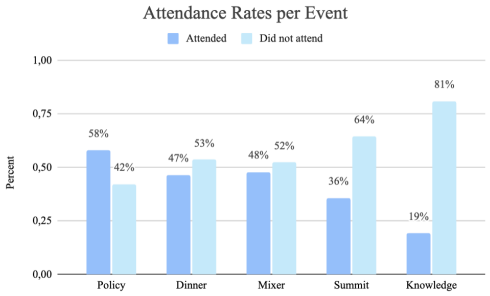
[1] "Om Swedenbio," SwedenBIO, https://swedenbio.se/om-swedenbio/ (accessed Mar. 7, 2024).
[2] "Swedenbio's Privacy Policy," SwedenBIO, https://swedenbio.se/en/about-swedenbio/swedenbios-privacy-policy/ (accessed Mar. 7, 2024).
[3] D. Jurafsky and J. H. Martin, "Speech and Language Processing", 2023
[4] J. Boye, "DD1418: 8a: More on classification, 8b: Logistic regression," in Canvas, DD1418 HT23 Språkteknologi med introduktion till maskininlärning (51872)
[5] "What is Random Forest?," IBM, https://www.ibm.com/se-en/topics/random-forest (accessed Mar. 6, 2024).
[6] "What is XGBoost?," NVIDIA Data Science Glossary, https://www.nvidia.com/en-us/glossary/xgboost/ (accessed Mar. 6, 2024).
[7] Shung, K.P. (2020) Accuracy, precision, recall or F1?, Towards Data Science. Available at: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 (Accessed: 26 April 2024).
[8] "Understanding Baseline Models in Machine Learning", Medium, https://medium.com/@preethi_prakash/understanding-baseline-models-in-machine-learning-3ed94f03d645 (accessed Apr. 20, 2024).
[9] N. Barney and J. Biscobing, "What is customer engagement?: Definition from TechTarget," Customer Experience, https://www.techtarget.com/searchcustomerexperience/definition/customer-engagement (accessed Mar. 6, 2024).
[10] W. M. Lim, T. Rasul, S. Kumar, and M. Ala, "Past, present, and future of customer engagement," Journal of Business Research, vol. 140, pp. 439–458, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296321008213
[11] T. Chavez, C. O'Hara, and V. Vaidya, Data driven: harnessing data and AI to reinvent customer engagement. New York: McGraw-Hill Education, 2019.
[12] Jeremy K. Nguyen, Adam Karg, Abbas Valadkhani and Heath McDonald (2022) Predicting individual event attendance with machine learning: a 'step-forward' approach, Applied Economics, 54:27, 3138-3153, DOI: 10.1080/00036846.2021.2003747
[13] O. J. Leong and R. Mafas, "A REVIEW ON THE PREDICTIVE ALGORITHMS FOR EVENT ATTENDANCE PREDICTION IN EVENT-BASED SOCIAL NETWORKS", Journal of Critical Reviews, July 2020
[14] J. Awad, "Event Attendance Prediction and Recommendation System: A Comparative Study of Machine Learning Algorithms", 2023-24

APPENDIX I: EVENT-DATA

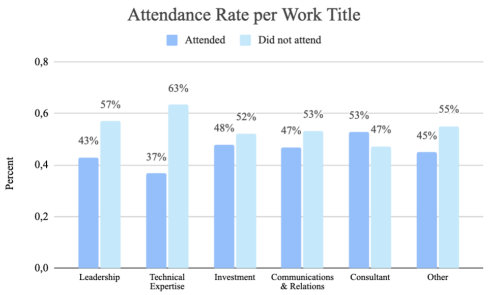


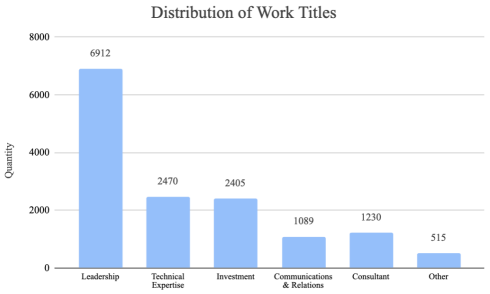


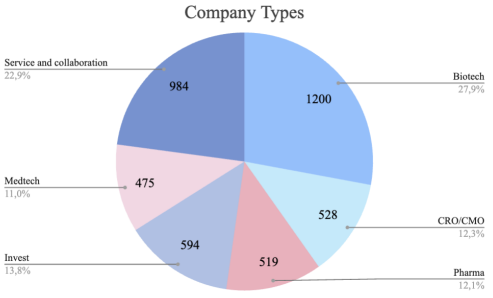


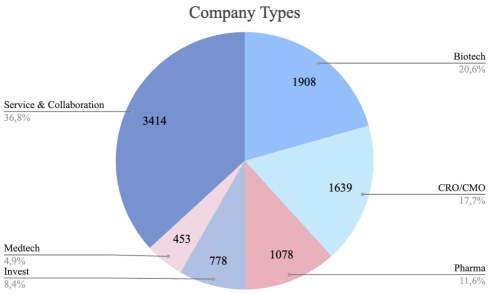


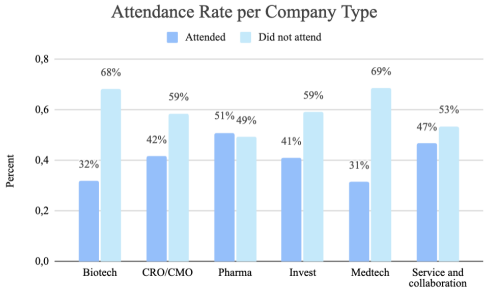
APPENDIX II: EXPO-DATA

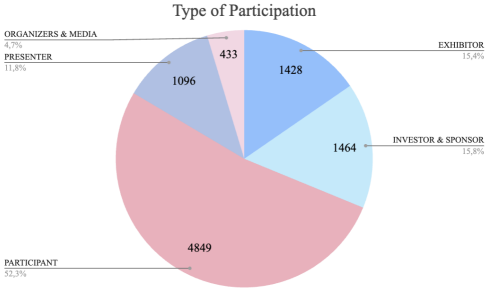












TRITA-EECS-EX-2024:253
Stockholm, Sverige 2024

www.kth.se