DATA603-24X

RESEARCH ESSAY

SHAHRON SHAJI
51781227

CONTENTS

INTRODUCTION

During my internship at Data Consulting, I had the incredible opportunity to work on a project predicting gym membership conversion rates using integrated datasets. Our focus was on Genesis Gym Ballarat in Victoria, Australia, and we aimed to develop a predictive model using mobile location data from Azira, Census data from the Australian Bureau of Statistics (ABS), and the gym's first-party membership data.

I was excited to apply my data science skills to a real-world problem, knowing that working with large, complex datasets would be both challenging and rewarding. My main task was to explore the data, identify relationships between key variables, and determine if these patterns could help predict membership conversions. While the project had its complexities, it was a fantastic learning experience—giving me hands-on exposure to real-world data challenges, problem-solving, and the impact of data-driven decision-making for businesses.

ORGANISATION

Data Consulting is all about helping businesses make sense of data by giving them easy access to third-party datasets and advanced technological tools. Their mission is straightforward but impactful: to make data-driven decision-making accessible, practical, and affordable for businesses of all sizes.

What sets them apart is their understanding that not every business has the time, resources, or expertise to handle data on its own. Instead of a one-size-fits-all approach, they customize their services to meet each client's specific needs. Whether it's a small business trying to understand its customers better or a large corporation looking for in-depth geospatial insights, Data Consulting provides tailored solutions to fit their goals.

One of their standout offerings is aggregated datasets that help businesses analyze visitor behavior, track movement patterns, and connect demographic data to specific locations using Statistical Area Level 1 (SA1) codes. This allows businesses to identify trends and make smarter, more targeted decisions. For example, understanding how often people visit a location and how long they stay can help businesses fine-tune their marketing strategies, enhance customer engagement, and even predict future trends.

A big focus for Data Consulting is human movement data—insights into how people move through physical spaces, where they go, how frequently they visit, and the patterns that emerge over time. This information is invaluable for businesses across industries, from retail to urban planning. For instance, a restaurant might notice a surge of visitors from a particular demographic during lunchtime and adjust its hours or promotions to better cater to them.

What really stood out to me during my time at Data Consulting was their collaborative approach. They don't just hand over raw data and leave clients to figure it out—they work alongside them to ensure the data is relevant, easy to interpret, and secure. This hands-on guidance helps businesses transform complex datasets into clear, actionable insights. Plus, they place a strong emphasis on ethical data management, ensuring all sensitive information is handled responsibly and in line with privacy regulations.

By offering scalable, project-based solutions, Data Consulting helps businesses stay ahead in a constantly evolving market. Their expertise in data integration and visualization empowers companies to make smarter decisions and optimize their operations. What makes them truly stand out is their ability to take massive, complex datasets and turn them into useful, practical tools that businesses can actually use to grow and succeed.

PROBLEM STATEMENT

The goal of this project was to see if we could predict gym membership conversion rates by combining mobile location data, first-party membership data, and ABS Census data. In simple terms, we wanted to understand how population movement and demographic characteristics could help Genesis Gym Ballarat improve its marketing and operations.

To make this possible, we had to align these different datasets geographically. We used Statistical Area Level 1 (SA1) codes to connect mobile movement data with detailed demographic information. These codes gave us a standardized way to analyze location-based trends, ensuring our findings were both accurate and meaningful.

One of the biggest challenges was figuring out whether mobile location data could reliably reflect broader demographic trends. This required a structured approach, including data integration, hypothesis testing, and predictive modeling. We had to clean and merge datasets to ensure consistency before analyzing patterns and drawing conclusions.

Having a clear problem statement was crucial—it kept us focused and helped us navigate the complexities of the data. Our ultimate goal was to give Genesis Gym Ballarat actionable insights by identifying key factors influencing gym membership conversions. By developing predictive models, we aimed to help them better target potential members and refine their marketing strategies.

But this project was about more than just numbers and algorithms. It was about understanding human behavior—where people live, where they go, and what drives their decisions. By tapping into these patterns, we hoped to give Genesis Gym Ballarat practical insights that would help them connect with their community and grow their business.

GOALS

This project had a few key goals that shaped our entire approach, from preparing the data to building the predictive model. The main objective was to analyze correlations and develop a model that could predict future gym membership conversion rates based on socio-economic and demographic factors.

A crucial part of this process was integrating different datasets—combining first-party membership data with CEL data and ABS Census data. To do this, we used SA1 geographic codes to map individuals to specific locations, ensuring all the data was aligned properly. This step was essential because maintaining geographic consistency helped us build a strong foundation for accurate analysis and reliable insights.

BACKGROUND

For organizations looking to improve their services and engage users more effectively, understanding demographic and behavioral patterns is key. In this project, we brought together multiple datasets—mobile location data from Common Evening Locations (CEL) and Common Day Locations (CDL) reports, first-party membership data, and census files—to analyze regional demographics and movement behaviors. Our goal was to provide Genesis Gym Ballarat with valuable insights to support their business decisions.

Throughout the process, we focused on maintaining data quality and consistency, ensuring that our analysis was both accurate and reliable. Since we were working with sensitive demographic and behavioral information, we took a careful and ethical approach to handling the data. More than just combining datasets, our aim was to extract meaningful insights that could help the gym refine its strategies and better connect with its community.

DATA SOURCES

The project involved working with three core datasets, each bringing a unique perspective to the analysis:

- **Census Data (AUS_SA1_G files)**: The census data consisted of seven different tables, each providing insights into key aspects of the local population. These tables covered age distribution, gender breakdowns, languages spoken at home, health conditions, household income levels, and housing types. Together, they helped us build a detailed picture of the community at the Statistical Area 1 (SA1) level. This dataset served as the foundation for understanding the demographics of the region and was crucial for making meaningful connections between population characteristics and gym membership trends.
- **Common Evening Locations (CEL) Data/ Common Daytime Locations (CDL) Data**: Also known as Azira data, this dataset provided mobile movement insights, including hashed user IDs, SA1 codes, assumed home locations (latitude and longitude), and visit frequency to specific locations. This data was essential for understanding foot traffic patterns—how people moved through different areas and how often they visited certain locations.
- **First-Party Membership Data**: This dataset contained details about Genesis Gym Ballarat's members, including their age, gender, and home location. This information was the key to identifying the characteristics of existing gym members and predicting potential new memberships.

To ensure a meaningful analysis, we prioritized variables that could offer actionable insights. Foot traffic patterns from the CEL/CDL data (visitor counts and visit durations) helped us understand movement trends, while census data (age, income, household composition) provided valuable context about the people contributing to those trends. The gym's first-party data allowed us to directly analyze member characteristics. By linking these datasets, we aimed to uncover patterns that could help Genesis Gym Ballarat refine its marketing and business strategies.

The first challenge was making sure all three datasets were properly aligned. To do this, we used SA1 codes as a common geographic reference, meticulously mapping and validating the data to maintain accuracy. Before diving into the analysis, we also focused on data cleaning to eliminate errors and inconsistencies, ensuring that our insights would be reliable and precise.

DATA PREPARATION AND LITERATURE REVIEW

Before diving into the technical work, we took a step back to conduct a thorough literature review. Our goal was to understand how mobile movement data had been used in similar research projects and to gather insights that could shape our approach. This helped us set realistic expectations and provided valuable context for our analysis.

We explored studies on geospatial analytics, predictive modeling with demographic data, and the business applications of movement data. These studies highlighted how combining spatial analysis with demographic insights could reveal meaningful patterns in visitor behavior. Seeing real-world applications of geospatial data helped us appreciate how our project could provide practical value to businesses like Genesis Gym Ballarat.

The time spent on the literature review proved to be incredibly valuable. It gave us a strong theoretical foundation and allowed us to approach the project with clarity and confidence. Learning from past research helped us refine our methodology and ensure we followed best practices. Most importantly, the insights we gained helped us see how our findings could translate into actionable recommendations, turning our analysis into a useful tool for business decision-making.

DATA CLEANING AND MERGING PROCESS

To begin our analysis, we first merged all Census datasets using SA1 codes and cleaned them by removing any missing or null values. We also ensured that data types were correctly formatted, keeping the dataset well-structured and ready for analysis.

The next step was to map the latitude and longitude of the first-party membership data to SA1 codes within the CEL dataset. We accomplished this using QGIS, a geographic information system that allowed us to accurately link gym members' home locations to specific SA1 regions.

Once we successfully mapped the first-party data, we refined the dataset further to prepare it for correlation analysis. These refinements were essential for ensuring that our data was comprehensive, clean, and meaningful.

ENHANCING THE DATASET WITH KEY METRICS

1. **Distance Traveled**
    a. We merged the *Average Distance in Miles* column from the CEL dataset with our first-party data using SA1 codes.
    b. This gave us insights into how far members were traveling to visit the gym, which could help in understanding conversion rates and visitor behavior.
2. **Income Brackets**
    a. Using Census data, we identified the most common income ranges within each SA1 code.
    b. We added two new columns representing the lower and upper limits of these income brackets to analyze whether income levels influenced gym membership conversion.
3. **Birthplace Index**
    a. We calculated the probability of individuals being born overseas within each SA1 code.
    b. This allowed us to explore whether cultural diversity in a region had any impact on gym membership trends.
4. **Education Index**
    a. We analyzed the number of people in each age group attending educational institutions.
    b. By calculating the probability of educational attendance relative to the total population within each SA1 code, we could examine whether education levels correlated with gym visitation patterns.
5. **Age Distribution**
    a. We categorized individuals into age brackets (e.g., 0–19, 20–44, 45–64, 65+).
    b. We then calculated the percentage of each age group relative to the total population within their SA1 code to understand which age groups contributed more to gym visitation.
6. **Gender Distribution**
    a. We calculated the percentage of males and females within each SA1 code.
    b. This helped us analyze whether gym visitation patterns were influenced by gender dynamics in different regions.

Each of these additional metrics enriched our dataset, allowing us to layer demographic and behavioral insights for a more comprehensive analysis.These refinements ensured that our analysis would be grounded in real-world context and that our findings would offer meaningful insights for Genesis Gym Ballarat.

We developed three types of conversion rates:

1. **First Party Conversion Rate**: The proportion of individuals from each SA1 code in the first-party dataset relative to the total number of 18+ individuals in the same SA1 code based on Census data.

2. **CEL Conversion Rate**: The proportion of hashed Ubermedia IDs in each SA1 code relative to the total number of 18+ individuals in the corresponding SA1 code from Census data.

3. **CEL to First Party Conversion Rate**: The proportion of individuals from each SA1 code in the first-party dataset relative to the count of hashed Ubermedia IDs in the same SA1 code.

After evaluating these rates, we selected the **First-Party Conversion Rate** as the most reliable metric for our analysis. While the **CEL Conversion Rate** helped understand the likelihood of potential casual visitors and members in specific regions, the **CEL to First-Party Conversion Rate** produced skewed results, highlighting limitations in the Azira dataset for calculating this metric accurately.

CORRELATION ANALYSIS AND HYPOTHESIS TESTING

To understand how different factors influence gym membership conversion rates, we used **Pearson correlation**. We chose this method because it's simple, effective, and helps measure the strength of the relationship between two continuous variables. We started by creating a **correlation matrix and a heatmap**, which allowed us to visualize how key variables—such as age, income level, education index, birthplace index, and travel distance—were related to the conversion rate. This helped us identify potential trends and patterns in the data.

Based on our initial insights, we developed multiple hypotheses to test the predictive model. Our final **null hypothesis** stated:

"Common lower income, education index, birthplace index, and distance in miles do not have a significant effect on the gym membership conversion rate."

After running our analysis, we observed a **low F-statistic value**, which allowed us to reject the null hypothesis. This finding confirmed that at least one of these socioeconomic factors, along with travel distance, does have a significant impact on the conversion rate. By identifying these key influences, we gained valuable insights into the factors that drive gym membership conversions, helping Genesis Gym Ballarat refine its marketing and outreach strategies.

MODEL DEVELOPMENT AND TESTING

To accurately predict gym membership conversion rates, we explored several modeling techniques and ultimately selected the **Gradient Boosting Regressor** as our final model. This choice was based on its strong ability to capture complex patterns in the data. To prevent overfitting, we introduced a small amount of noise into the dataset, ensuring that the model could generalize well to new inputs.

Additionally, we conducted a **Variance Inflation Factor (VIF) analysis** to check for multicollinearity between predictor variables. The results confirmed that there were no significant collinearity issues, reinforcing the reliability of our model.

Our final model delivered impressive performance metrics, achieving an **R-squared (R^2) value of 0.881**, indicating strong predictive power, and a **Root Mean Square Error (RMSE) of 0.007**, suggesting minimal prediction errors. To further validate the model's effectiveness, we tested it on a new dataset from the South Australian Aquatic and Leisure Centre (SAALC). Using the same methodology, we generated demographic and behavioral indexes for each SA1 code and predicted the corresponding conversion rates. The predictions were then reviewed by Karen, our industry mentor, who confirmed their accuracy with **80% confidence**. This validation reinforced the model's effectiveness, demonstrating that it was not only applicable to Genesis Gym Ballarat but could also be used to predict potential gym memberships for other fitness centers.

PROFESSIONAL DEVELOPMENT AND WORKPLACE EXPERIENCE

The project began with a common meeting where James provided an overview of the timelines, rules, and regulations. This session was crucial in helping me understand the project's overall structure and expected milestones. Following this, we had an in-person meeting with Ricki, our industry partner, who gave us a clear explanation of the dataset and his expectations. This meeting helped clarify many of our initial doubts and provided direction for our analysis. We then reached out to Karen, who manages the data on the industry side. Her deep understanding of the various datasets proved invaluable in ensuring we had the right information for our analysis.

Throughout the project, we had regular weekly meetings with Zac, our university mentor. These sessions provided consistent feedback, ensuring that we stayed on track and made the necessary improvements along the way. After developing our model, we had a meeting with James, and his valuable insights helped us refine the model further, leading to improved results. The feedback from the presentation and poster also played a significant role in enhancing our work, allowing us to make key adjustments and ultimately deliver a better final output. Both our industry partners and university mentors were incredibly supportive, guiding us through challenges and helping us achieve better results. Working on this project allowed me to collaborate closely with supervisors and mentors, significantly enhancing my communication skills. Additionally, I gained hands-on experience with various tools like QGIS, which expanded my technical knowledge and analytical capabilities.

There were, however, some challenges along the way. A delay in receiving the dataset initially meant we couldn't start the analysis on time. Later, we encountered server issues with MADS, which temporarily halted our work. While we couldn't always hold daily meetings, we effectively coordinated through Microsoft Teams to stay aligned and collaborate efficiently. Through a series of discussions and presentations, I became more confident in communicating my ideas, and the experience significantly improved my public speaking and presentation skills.

Once we had all our meetings done with our industrial supervisors, we had a clear idea about what to do and what to achieve with out limited data. We figured out different methods to approach this and interestingly there was no correlation we could see and look deeper into. But we didn’t give up here, logically speaking the distance, income, ethnicity and education should play a good role in figuring out the number of people converting. So I suggested we would do a hypothesis testing to see the significance of these variables with conversion rate and we were on the right track. This gave us the confidence to move forward with the modelling and interesting we were able to finally bring up a model with an accuracy above 80 percent which we could test it with and confirm with the industrial supervisors.

Overall, this project has been a transformative experience, helping me grow both professionally and personally. It not only strengthened my technical expertise and problem-solving abilities but also reinforced my confidence in teamwork, collaboration, and effective communication.

PERSONAL GROWTH

The internship experience at Data Consulting has fostered significant personal and professional growth. Initially, challenges arose in integrating large datasets, demanding problem-solving strategies and the application of advanced data science techniques. Overcoming these challenges emphasized the value of adaptability and collaboration in a professional setting. Moreover, the ethical considerations involved in handling sensitive demographic and behavioral data were highlighted, ensuring responsible data usage throughout the project. This study provided essential insights into data integration challenges, merging heterogeneous data sources, and applying predictive models for actionable insights, thereby emphasizing real-world applications of data science methodologies.

Beyond technical skills, the mentorship I received from Karen was instrumental in my learning process. She provided guidance not only on technical tasks but also on navigating workplace dynamics and managing time effectively.

CAREER GOALS

While I enjoy and find the role of a data scientist fascinating, my true aspiration, as someone with a strong programming background, is to become a data engineer in the future. This doesn’t necessarily mean I am not inspired to become a data scientist in the future, but I have no intention to leave behind the raw coding experiences I acquired throughout the course of this career and my professional journey as a software developer. I actively work on coding and building pipelines as part of my personal projects, which not only fuels my passion but also enhances my ability to handle the coding aspects of programs. This aspiration makes navigating and managing the coding process much easier and more efficient.

SKILLS DEVELOPMENT

I gained new skills, including the use of QGIS, but what truly captivated me was interpreting the data we had and developing solutions to meet specific needs. The insights and ideas shared by my research team inspired me and introduced new perspectives on how to approach and interpret data effectively. Another interesting idea that captivated my mind was the intentional introduction of noise into the data that would help the model to finally understand the pattern and give an better accuracy.

ROLE OF GENERATIVE AI

Throughout this project, I used ChatGPT as a supportive tool to enhance my work and streamline complex tasks. It helped me troubleshoot Python code and tackle challenges with dataset merging and feature engineering, such as calculating demographic percentages and conversion rates. ChatGPT also clarified complex machine learning concepts, like Gradient Boosting and Elastic Net, making it easier for me to apply these models effectively. While it was a helpful resource for guidance and problem-solving, all the analysis, modeling, and interpretations reflect my own understanding and effort. I've acknowledged its use to ensure transparency and maintain academic integrity.

INDUSTRY INSIGHTS

The internship provided valuable insights into how businesses leverage data to make decisions. I learned that:

- **Data privacy is paramount**, especially when dealing with personal information.
- **Data integration and cleaning is a complex but essential task** in ensuring accurate analysis.
- **Predictive models can drive actionable insights**, improving decision-making processes.

These insights shaped my understanding of data science in practice and highlighted the importance of ethical data handling.

CONCLUSIONS

The most valuable aspect of this internship was the hands-on experience with real-world data challenges. I learned to navigate technical obstacles, collaborate effectively, and apply advanced data science techniques to drive meaningful insights.

The experience exceeded my initial expectations by providing opportunities for both technical and professional growth. I gained a deeper understanding of data science applications in the industry and developed skills that will be invaluable in my future career.

For future programs, I suggest providing datasets earlier in the process to allow more time for exploration and offering workshops on specific tools like QGIS to accelerate the learning curve. Overall, this internship was a transformative experience, equipping me with practical skills, industry insights, and a renewed passion for data science.

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to everyone who contributed to the success of this project. First and foremost, I would like to thank **Data Consulting** for providing me with this invaluable opportunity to work on a real-world data science problem. Their expertise, resources, and guidance played a crucial role in shaping my understanding of data integration, predictive modelling, and business applications of data science.

A special thanks to **Ricki**, our industry partner, for his insightful feedback and for helping us understand the dataset and its real-world significance. His guidance enabled us to approach the project with clarity and purpose. I would also like to express my appreciation to **Karen**, whose mentorship was instrumental throughout the project. Her deep knowledge of the datasets and her willingness to share insights greatly enriched my learning experience. Her constructive feedback helped refine our approach and improve our model's accuracy.

I am also grateful to **James** for his leadership and for setting clear expectations at the beginning of the project. His insights and feedback during our model development phase were invaluable in fine-tuning our approach. A heartfelt thank you to **Zac**, our university mentor, for his continuous support and regular feedback. His expertise in data science and predictive modelling provided us with a strong technical foundation, ensuring that our approach was methodologically sound.

Lastly, I would like to acknowledge my team members for their hard work, collaboration, and dedication throughout the project. Working alongside them has been an enriching experience, and their contributions played a vital role in our collective success.

This experience has been a transformative journey, and I am deeply appreciative of the opportunity to learn, grow, and apply my skills in a real-world setting.

REFERENCES

1. **Gradient Boosting Regressor**
   Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
2. **Machine Learning Theory and Applications**
   Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
3. **Data Preprocessing and Feature Engineering**
   Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
4. **Scikit-Learn Documentation**
   Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://scikit-learn.org/stable/

5. **Python for Data Science**
   McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
6. **Business Applications of Predictive Modeling**
   Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.
7. **Online Sources for Industry-Specific Insights**
   - Kaggle Datasets: https://www.kaggle.com/
   - Towards Data Science Blog: https://towardsdatascience.com/
   - Google Scholar: https://scholar.google.com/