DBA4714

Deep Learning and Generative AI in Business

Group Project

Ground Zero

Name	**Matriculation Number**
Andreas Lukita	A0221743M
Yu Bixun	A0218248B
Yu Tianyi	A0218288U

Content Page

1. Business Problem and Motivation

In the fiercely competitive hospitality industry, profit-driven entities such as hotel management continually seek innovative strategies to maximise profits and optimise pricing strategies to stay ahead of market fluctuations in order to ensure long-term business growth and expansion. This could be achieved with revenue maximisation and cost minimisation.

However, the inherent uncertainty surrounding booking cancellations poses a significant challenge for hotels striving to maximise their profit potential. According to a study conducted by D-Edge Hospitality Solutions, the average global booking cancellation rate in the hospitality sector increased by 7.1% year-on-year from 2014 to 2018, with the current global cancellation rate at around 40% (D-Edge, 2019). Analogous to the newsvendor problem in operation management, this situation is undesirable for the hotel management as cancelled reservation would mean missed revenue opportunities and underutilization of hotel resources because the unused rooms represent potential income that cannot be recovered.

Accounting for the emerging information and dataset containing valuable information on booking reservations and their outcome, the hotel faces the pressing need to develop effective revenue management strategies. One strategy is to maximise hotel occupancy rate and this could be achieved by performing certain levels of overbooking. Leveraging advanced predictive analytics techniques, such as machine learning (ML) algorithms, the hotel seeks to anticipate whether customers are likely to honour their reservations or cancel them before their arrival day.

However, overbooking strategy is a double-edged sword — while it can potentially help to maximise revenue, overdoing it can lead to customer dissatisfaction in the event when we mispredict customer arrival. Take for example, if the model predicts a cancellation coming from customer A, the hotel allows the room to be booked by a new customer B in hopes of utilising the room for revenue. However, in the case of a false-positive scenario in which customer A turns up, the hotel will face consequences of having to compensate customer A by relocating him to another nearby hotel and potentially facing reputational damage.

Therefore, the hotel must navigate this trade-off judiciously by leveraging data-driven insights. By addressing these challenges head-on and integrating predictive analytics into their revenue management approach, the hotel aims to enhance revenue performance and improve operational efficiency, without compromising guest experience.

However, as no model is a perfect model, it is to be expected that there will be cases where predictions could go wrong. In such a scenario, we need to carefully weigh the cost associated with the wrong prediction. As such, the hotel needs to perform a cost-benefit analysis between the cost of wrong prediction of cancellation (false-positive) and the cost of wrong prediction of non-cancellations (false-negative).

2. Methodology

With the rise of Large Language Models (LLM), usage of Artificial Intelligence (AI) can be purposed for a wide variety of applications ranging from informational searches to data analytics. According to a Harvard Business Review article by Waber & Fast, usage of Generative AI and LLMs can potentially grow businesses by US$4.4 trillion globally every year. In light of this, the group wishes to explore the effective usage of LLMs such as GPT4 (ChatGPT), Gemini Advanced and Claude in terms of business and data analysis. Specifically, where feasible, the group uses LLMs to brainstorm ideas, discuss insights, and generate code. The group did not intend to prove that AI is able to replace and take over human roles and functions as data analysts. Instead, the group aims to demonstrate how AI and humans can collaborate and uplift each other's work to produce better results.

This report aims to focus on the following 3 aspects:

1. Solving the business problem of hotel's revenue management
2. Integration of LLMs in solving the business problem
3. Analysing, evaluating, and critiquing the outputs produced by the 3 LLMs

The group obtained the [Hotel Reservations Dataset] through Kaggle (Raza, 2023). In the initial stage, the group experiments with different LLMs to get a sense of the respective output. Thereafter, we compile the best and most relevant outputs by the LLMs to form a list of consolidated prompts. In order to prompt in a specific direction to reach the same end goal across all the LLMs, we decided to guide the LLMs using the consolidated prompts in the following way — using Primary and Secondary questions. In short, Primary questions are mandatory questions that should be prompted to all the LLMs to ensure the same basis of comparison. Meanwhile, Secondary questions are follow-up questions specific to the output of the respective LLMs to capture the different nuances of the respective LLMs.

We approach the business problem by first performing descriptive analytics to better understand the dataset. Subsequently, we built a binary classification model to predict the likelihood of a customer cancelling their reservation before their actual arrival day. The model is then tuned to suit the risk appetite of the business stakeholders, and the cost savings is calculated through the combination of different prediction results using a Confusion Matrix. At the end, the group intends to convince the business stakeholders to adopt the AI model as it generates value for the hotel management in terms of cost minimisation.

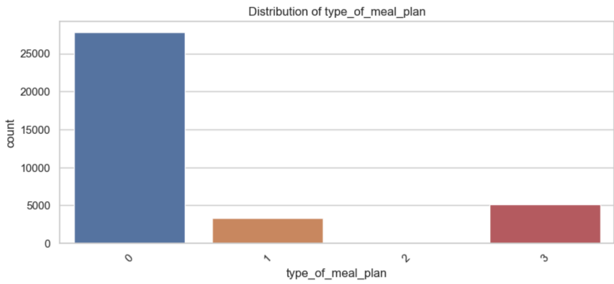
The group has attached notebooks containing code that each LLM generated (with little to no human assistance). Thereafter, we created a separate notebook compiling the best output from each of the LLM, and added human modification on top to value-add and improve the final result. As a wrapper to present the work produced, the group has created a Streamlit Interface to allow users to have a better understanding of the result during the presentation.

3. Analysis on output of LLMs

a. Descriptive analytics

Generally, all the LLMs were able to provide a comprehensive passage of information with regards to the group's primary and secondary prompts. For instance, the prompts for finding various business problems or generating related problem statements were answered with logical and feasible outputs (Refer to Appendix 7 A). Certain outputs were backed by citations given by the LLMs, such as statistics or facts from news articles or websites (Refer to Appendix 7 A5.).

Moving on to descriptive analysis, all 3 LLMs were able to produce adequate codes to perform data quality checks on the dataset. Ranging from checking for duplicated data to missing values, the code provided could work and run on the group's IDEs (Integrated Development Environment).



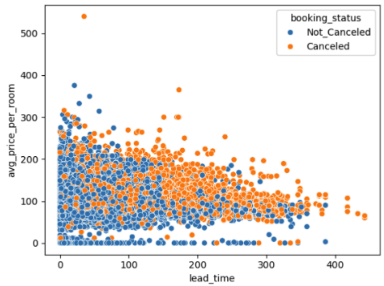
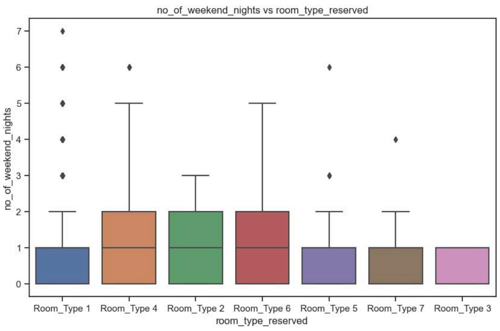


Diagram 1 & 2: Bar Chart showing counts of meal plans and Scatterplot showing relationship between average price per room and lead time colour-coded by booking_status.

Some of the data visualisation charts produced by the LLMs on univariate and bivariate analysis were adequate. With reference to the bar chart coded by ChatGPT 4.0 above (Diagram 1 on the left), the usage of such a chart to show the number of counts of data for each "type_of_meal_plan" is appropriate as it clearly shows the amount of data points for each feature. As for the scatter plot chart coded by Gemini (Diagram 2 on the right), it provides us the distribution of average price per room against the lead time. The scatterplot (Diagram 2 on the right) can aid the group in discovering any potential relationships between the 2 numerical variables, and any changes in such patterns with respect to booking status.



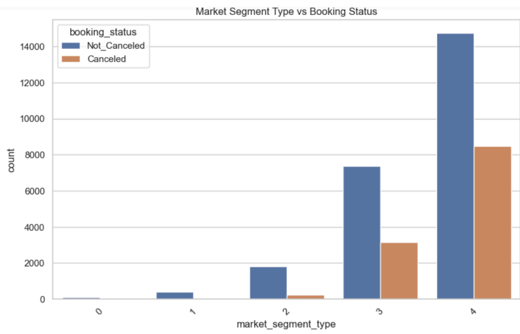


Diagram 3 & 4: Box-&-Whisker Plot on distribution of no. of weekend nights across room types and Bar chart showing counts of cancelled vs not cancelled rooms for market segments.

On the other hand, there are some charts produced by the LLMs that were not the best way to visualise such data. For example, the box-&-whisker plot above (Diagram 3 on the left) is not the best way to visualise distribution of discrete numerical data like the number of weekend nights. A better alternative would be usage of histogram or bar charts to visualise such data. Additionally, the graph above on the right side shows a side-by-side vertical bar chart

(Diagram 4 on the right) showing the relationship between cancelled and non-cancelled bookings for each market segment. However, it would be more intuitive to use a normalised stacked bar chart to perform analysis on the proportion of booking statuses across the market segments to achieve more meaningful analysis. In essence, the LLMs recommend and follow the basic procedure required to perform descriptive analysis tasks.

In a nutshell, though LLMs can support users by generating quick visualisations, a user's professional judgement and domain knowledge are still needed in validating the data and selecting the best graph or performing analysis to derive insights from the visualisations. Ideally, users should use directional stimuli, few-shots or instruction prompting to achieve their desired visualisations for univariate/bivariate analysis.

b. Predictive analytics

i. Feature Engineering

With instruction prompting, the LLMs were capable of dropping unwanted or potentially invalid data, as well as performing feature engineering on the dataset such as binning, feature creation or performing categorical encodings (Refer to Appendix 7 C). With zero-shot prompting, LLMs were also capable of distinguishing between different types of categorical encodings and their respective use cases. Feature scaling such as standardisation were also coded appropriately by the 3 LLMs to ensure efficiency in training models (Refer to Appendix 7 C6.). LLMs are also capable of creating and recommending new features based on the business problem. For example, LLMs can create a new feature that indicates a boolean data on whether the customer is staying alone. Hence, LLMs generally have the capability of aiding users in feature engineering to facilitate predictive analytics.

Out of the 3 LLMs, ChatGPT 4.0 seems to have the issue of not using the appropriate feature transformation for certain features. For example, ChatGPT 4.0 seems to provide code to perform ordinal encoding for “type_of_meal_plan” and “market_segment_type”, which may not be ideal without further investigation as there may not be any form of logical ranking or hierarchy in meal plans or market segment types (Online booking vs Complementary booking).

ii. Performance of Predictive Models by LLMs

Base Model	GPT4	Gemini	Claude
Accuracy	0.76	0.81	0.76
Precision	0.69	0.84	0.78
Recall	0.48	0.89	0.90
F1-Score	0.56	0.87	0.84
AUC Score	0.69	0.77	0.79

Table 1: Performance of Logistic Regression Models (With hyperparameter tuning).

LLMs generally could distinguish and provide us with different forms of metrics to measure the performance of predictive models, ranging from F1-Score to ROC-AUC score. Depending on the nature of the business problem, LLMs is also capable of programming out different predictive models, ranging from Logistic Regression to XGBoost Classifier. This way, it can support prompt users with measuring the performance of different models. Therefore, by utilising directional stimulus prompting, users can potentially use LLMs to derive the appropriate models for their business problem and get the best-performing model by measuring their performance metrics.

Generally, we observe that for the base model of comparison (Logistic Regression), the result produced by Gemini is on average better than GPT and Claude. The group does acknowledge that rigorous scientific experimentation has not been performed to validate whether this observation is consistent or purely due to chance. *(Please refer to attached notebooks to view the performances of other classification models if needed.)*

To derive the best model generated by our respective LLMs, we utilise directional stimulus prompting, which guides the conversation towards exploring alternative models such as XGBoost and SVC after completing work on the logistic regression model. This technique directs our focus towards the next steps or tasks to be addressed, allowing for a more comprehensive comparison among different ML models. By employing this approach, we aim to identify the optimal model for our dataset, considering factors such as performance metrics, model complexity, and generalisation capabilities. This systematic exploration helps in making informed decisions and selecting the most suitable model for our specific ML task.

4. **Improvised Code to tackle business problem**

From all the outputs produced by the 3 LLMs, the group selects the most relevant results and have manually collated the result into one final notebook. The group value-added to the LLM outputs by adding thought-processes that came with accumulated experience and exposure in the world of data analysis. These include writing proper code in classes, iteration in tuning hyperparameters, and deciding on probability decision threshold, etc.

Specifically, the hyperparameters to the model are tuned to minimise cases of false positives. The group wraps the 3 classification models into a soft voting classifier (XGBoost 0.5, Logistic Regression 0.3, and SVC 0.2) to obtain a result that is more robust to noise in the data. By default, the probability decision threshold for a classification model is set to be 0.50, and the common way of balancing between precision and recall is to use f1-score (harmonic mean of precision and recall). However, in our business context, since precision and recall are to be weighted differently, we cannot rely solely on f1-score as the primary metric. Instead, we opt to use F-beta score to adjust the proportion of importance of recall and precision. In short, we set beta to be 1.5, that is, recall is 1.5 times more important than precision. This aims to reduce the cases of false positives (predicted as cancelled but in reality otherwise), essentially improving our model performance as seen from the Confusion Matrix below where the number of false positives decreases from 494 to 142. *(Note that in our code, the* *labelling of the target variable is counterintuitive due to LabelEncoder labelling Cancelled as 0 and Non-Cancelled as 1. This is unlike the usual convention where minority classes (Cancelled) are often encoded as 1 instead of 0.)*

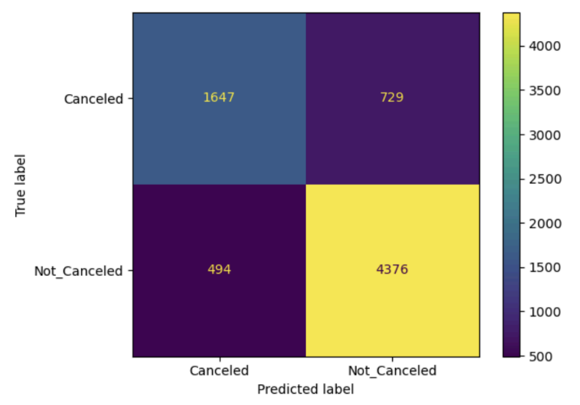


Diagram 5: Confusion Matrix before adjusting probability decision threshold.

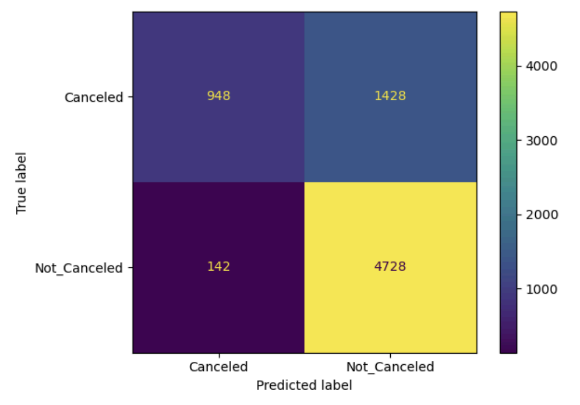


Diagram 6: Confusion Matrix after adjusting probability decision threshold.

a. Performance of Soft Voting Classifier (Human Assistance)

Metrics	**Value**
Accuracy	0.78
Precision	0.77
Recall	0.97
F1-Score	0.86
AUC Score	0.89

Table 2: Performance of Soft Voting Classifier (With human assistance).

5. Limitations of LLMs

a. Across all LLMs

Catching unusual data or specified analysis

According to the group's findings, all 3 LLMs were unable to identify the potentially odd occurrence of $0 for "average_price_per_room" given the business context (Refer to Appendix 7 B5.). While this is expected in Gemini and Claude as we only uploaded the first 20 rows of data to the interface limitation, ChatGPT is also unable to identify and point out the unusual occurrence of $0 for "average_price_per_room" even after performing instruction prompting (Refer to Appendices 7 B8., B9. and B10.). For the case of Gemini, few-shots prompting was performed, explicitly pointing out the data point to the LLM, which resulted in the LLM recognising the unusual nature of such data and giving possible reasons or implications behind them.

Another specific example would be that the 3 LLMs failed to detect and point out the presence of data rows with an invalid date feature such as 29 February 2018. In short, there are no 29 days in a February on non-leap years, and without Instruction Prompting to identify and remove such data, the code generated by the LLMs would not work. Even with few-shot prompting, the 3 LLMs were generally unable to correctly point out specific analysis on certain patterns or outliers in data visualisations most of the time, and only provided very general comments on what are certain patterns or anomalies to watch out for. We observe that the LLMs are generally "lazy" and provide the bare minimum as a solution for prompt users — providing only basic code templates, without doing sufficient data validation to check whether the data makes real business sense in the first place.

Data formatting/conversion

For all 3 LLMs, they came up with codes that are not the most suitable for data type conversion, especially when dealing with datetime format. In this case, *data['arrival_date'] = pd.to_datetime(data['arrival_date'])* did not successfully perform the conversion if there was no assistance by any human users due to invalid dates. Although an interesting insight is that ChatGPT 4.0 was able to recognise and re-perform data conversion upon errors, but ultimately failed to perform the conversion to our expected output.

Non-retention of context

The group noticed that LLMs generally struggle with "remembering" the current status of the programming project. For instance, after performing train_test_split to split the dataset into training and testing data, LLM continues to perform feature engineering on the original dataset instead of the splitted data, leading to the insidious data leakage issues. While the code still continues to run, the model performance is more often than not inflated, painting a wrong picture of the capability of the model. In another instance (in ChatGPT 4.0), the LLM performed Ordinal encoding on "lead_time_category" earlier on and create a newly encoded column "lead_time_category", but used One-Hot encoding on the "lead_time_category" when building the pipeline, disregarding its earlier decision of using Ordinal encoding.

Produced sub-optimal codes with lots of warnings

While the LLMs produce codes that generally run, it is trained on older documentation with codes that are already deprecated. As external packages often update, this normally gives rise to warnings, should the package version we are using differ from the version of which the LLM is trained on. Moreover, we observe that codes generated by Gemini were sourced and trained on Github repositories written by unknown users. While they produce the most commonly seen way of writing code, these may not be the best practices and could lead to multiple implications along the way *(i.e. not building models with Pipeline could accidentally lead to data leakage. Pipeline is essential to ensure the test set is unseen at all times)*.

b. <u>Specific to ChatGPT 4.0</u>

There were occurrences of misleading analysis on the data visualisation (Refer to Appendix 7 B4.) where most bookings were made for 2 adults but LLM stated that most involved "fewer" adults. Another interesting observation was that, despite ChatGPT 4.0 being capable enough to perform programming and analysis internally before giving users the output, it has a tendency to run into errors while performing such tasks halfway through the whole project (Refer to Appendix 1). As the group prompts further, the ChatGPT 4.0 revealed the error to be an internal issue in the execution environment utilised by ChatGPT 4.0 to perform the programming and analytics work (Refer to Appendix 2).

c. <u>Specific to Gemini Advanced</u>

Gemini Multimodal Model allows users to interact with and answer questions about data presented in text, images, video, and audio. The group tried this feature by uploading the images of plots generated using the code suggested by the LLM. The goal is to prompt the LLM to produce meaningful insights from the observations of the plot. Unfortunately, Gemini hallucinated and underperformed for the case shown below.

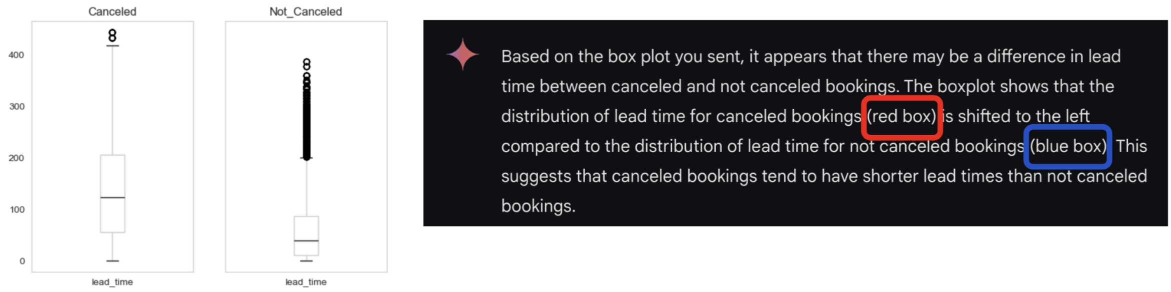


Diagram 7: Output by Gemini depicting inconsistent answers and references.

d. <u>Specific to Claude 3 Sonnet</u>

Despite cost considerations leading to the adoption of Sonnet instead of Opus, Claude 3 models have shown improved capabilities in performing analysis. However, there are constraints such as limitations on dataset size and time-based restrictions, as outlined in the provided reference (Refer to Appendix 3). Additionally, issues like overlapping variable names can impact model performance, such as leading to a recall of 1 for the support vector classification model when not training on the correct split data. Another observation is that the visualisations or graphical outputs produced by the Sonnet model may lack colour and additional information compared to other models, potentially affecting user experience and analytical insights. These factors highlight the need for careful consideration and potential adjustments in the modelling approach to address these challenges and optimise results.

6. Business Recommendation

a. Precision Recall Trade-off

In the revenue management in hotel business using the concept of optimising profits through overbooking and underbooking techniques, performance of models cannot solely rely on ROC-AUC Score. This is due to different magnitudes of cost with respect to different types of misclassifications. In this context, there might be different magnitudes of costs involved between overbooking and underbooking of rooms that need to be considered. Hence, a proper cost trade-off has to be calculated and analysed with metrics such as Recall or Precision.

At this point, it is essential to ask the questions on which costs are more painful to the company (i.e. whether the cost of failing to sell an unoccupied room hurts the bottom line more than the cost of relocation and bearing indefinite amounts of reputational damages).

b. Cost-Benefit Structure

To further guide decision-making, it is essential to quantify the cost trade-off between overbooking and underbooking. According to a Forbes report, hotel cancellation fees typically range from \$50 to \$300 or more, depending on the hotel's policies and length of stay (Peery Hotel, 2024). Additionally, the cost of overbooking can be substantial, ranging from \$152 (€142) (Zenderman, 2024). This cost includes expenses such as relocation costs, compensation, and potential negative impacts on the hotel's reputation and staff morale.

By considering these costs alongside precision and other metrics, we can make more informed decisions to optimise revenue management strategies. In our analysis, we assume the cost of underbooking to be equivalent to the mean daily rate of €100 derived from the dataset, reflecting the potential revenue loss from a unit unoccupied room. Conversely, the cost of overbooking is estimated to encompass not only a room upgrade from the current room type's average daily rate but also potential reputation damage incurred. These costs are not easily quantifiable and as such, the group has decided to put up an arbitrary figure of €350 cost per unit of overbooked room. This comprehensive approach ensures that both the direct and indirect costs associated with overbooking are accounted for, providing a more accurate representation of the financial impact on the hotel's operations.

c. Cost-Savings for the Hotel Business

Implementing our AI predictive model for hotel revenue management yields substantial cost savings and operational efficiencies — we mitigate the risks associated with false predictions, which incur significant expenses for the company. This allows the company to adopt the overbooking strategy with more confidence, instead of leaving the cancelled room as it is, or setting a random number of allowable rooms to be overbooked.

Take for discussion purposes the month of June 2018, in the absence of the predictive model, the total cost of managing these errors amounts to €129,100 per month. However, with the implementation of our AI model, the total cost was reduced significantly to €88,800 per month, resulting in a remarkable monthly cost savings of €40,300. Extrapolating this monthly saving over a year, assuming that the distribution of bookings for the other months are similar to that of June 2018 for convenience, our AI model has the potential to deliver annual cost savings of €483,600. These findings underscore the financial benefits and value proposition of integrating our AI predictive model into hotel revenue management strategies. Not only does it optimise revenue by minimising losses due to overbooking and underbooking, but it also enhances operational efficiency and profitability for the company.

While the analysis focuses on a single month to provide a snapshot of the potential cost savings achievable with the AI model, it is essential to note that analysis can be done on every month across one-year duration to have a more robust estimate of the cumulative impact on the annual company's bottom line. By continuously leveraging the predictive capabilities of the AI model, the company can sustainably optimise its revenue management practices and realise long-term financial benefits.

7. Future Business Recommendation

To optimise revenue management strategies further, it's recommended to delve into clustering techniques to categorise customers into more specific groups. By identifying distinct segments based on booking preferences and behaviour, we can tailor models to each subgroup, potentially enhancing prediction performance. Moreover, to bolster confidence in the AI model, we propose conducting Monte Carlo simulations to showcase its stability and effectiveness in the long term. Simulating booking cancellations will provide insights into the average additional revenue per month achievable with the proposed solution, aiding in convincing the business stakeholders of the value conferred by the model.

As hotel bookings could show trends and seasonal components potentially influencing the number of bookings across the year, it might be worthwhile to perform time-series analysis to forecast the number of bookings and cancellations in advance. This could aid us in deciding the maximum quota of allowable number of rooms to be overbooked.

Following this, model deployment is crucial, involving rigorous testing and implementation. A comprehensive profit analysis, comparing scenarios without and with the AI model, will quantify the impact on revenue and validate the investment. Finally, MLOps practices must be observed to maintain the model's performance over time.

8. Overall Conclusion

The purpose of the project is to firstly, to explore the benefits and limitations of using LLMs as a tool in tackling business problems, and secondly, to optimise revenue management in the hotel industry using ML techniques. By conducting experiments on LLMs with standardised set of primary questions and ad-hoc secondary questions, the group analysed and recognised the potential advantages of using LLMs in performing analysis in data science projects. In spite of LLMs being useful, there are certain limitations of LLMs that may run into risks of impeding the progress of solving the business problem if over-relied on. Being a double-edged sword, LLMs should be handled with understanding of how LLM works, effective prompting techniques as well as domain knowledge of the business problem. In essence, LLMs is not a one-stop solution for all kinds of business problems, but rather, a strong tool in aiding business analysts to tackle business problems with efficacy.

With respect to the hotel management business problem, LLMs certainly have its merits in contributing to a data-driven solution to ensure cost-savings. With hopes of future development in AI, qualities of outputs generated by LLMs may be expected and delivery of superior models can be performed efficiently.

9. References

Acharya, A. (2024, March 5). *Claude 3: Ai model suite: Introducing opus, sonnet, and haiku*. Encord. https://encord.com/blog/claude-3-explained/

D-Edge Hospitality Solutions. (2019, October 4). *How online hotel distribution is changing in Europe - D-edge*. https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/

Peery Hotel. (2024, March 10). *How much is a hotel cancellation fee? A comprehensive guide*. https://www.peeryhotel.com/how-much-is-a-hotel-cancellation-fee/

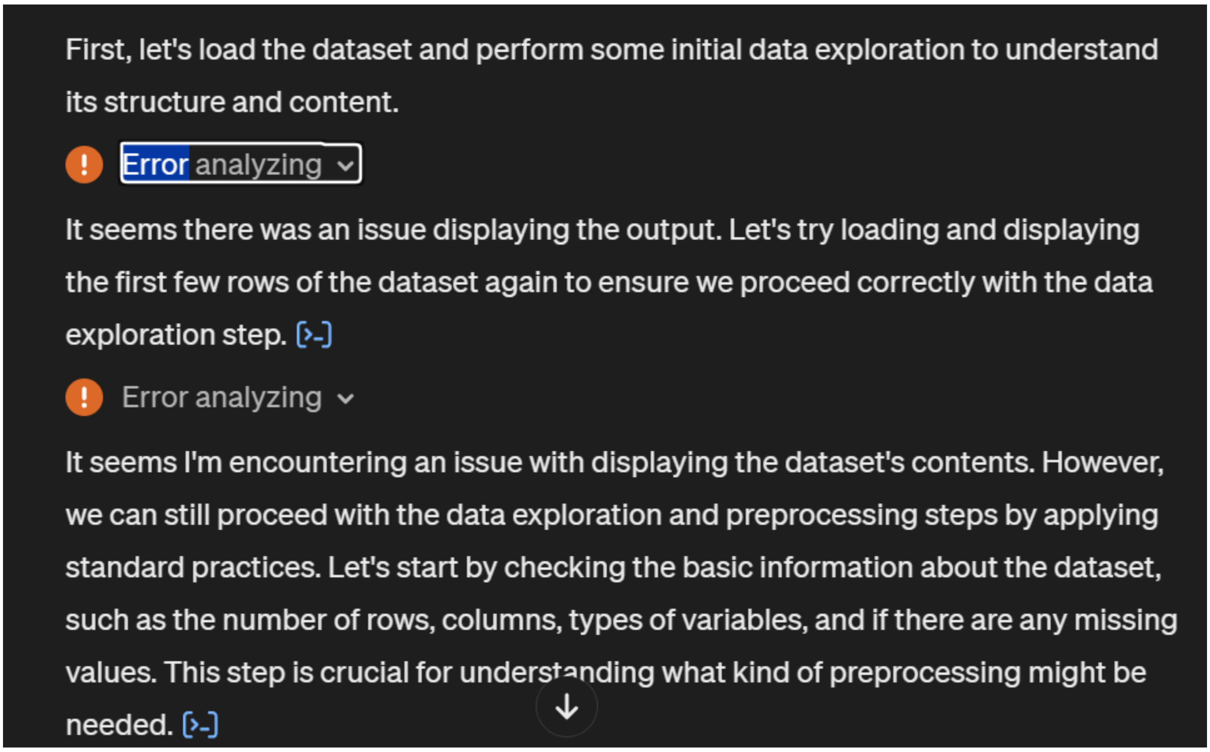
Raza, A. (2023, January 4). *Hotel Reservations dataset*. Kaggle. https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset

Waber, B., & Fast, N. J. (2024, January 8). *Is Genai's impact on productivity overblown?* Harvard Business Review. https://hbr.org/2024/01/is-genais-impact-on-productivity-overblown#:~:text=Large%20language%20models%20(LLMs)%20have,use%20of%20these%20same%20tools.

Zenderman, T. (2024, March 17). *Overbookings at your hostel: Cost breakdown*. BananaDesk Blog. https://bananadesk.com/blog/hostel-overbooking-cost-breakdown

10. Appendix

Appendix 1: Internal Error in running codes for ChatGPT 4.0



First, let's load the dataset and perform some initial data exploration to understand its structure and content.

Error analyzing

It seems there was an issue displaying the output. Let's try loading and displaying the first few rows of the dataset again to ensure we proceed correctly with the data exploration step. [>-]

Error analyzing

It seems I'm encountering an issue with displaying the dataset's contents. However, we can still proceed with the data exploration and preprocessing steps by applying standard practices. Let's start by checking the basic information about the dataset, such as the number of rows, columns, types of variables, and if there are any missing values. This step is crucial for understanding what kind of preprocessing might be needed. [>-]

Appendix 2: Explanation behind error in executing code in internal environment of ChatGPT 4.0

ChatGPT 4.0: The internal errors you've seen mentioned in my responses typically refer to technical limitations or issues within the environment where the code is being executed. These can happen for a variety of reasons, such as:

1. **Resource Limits**: There may be computational or memory limits that are exceeded when executing certain operations, such as training a complex machine learning model or handling large datasets.

2. **Tool Compatibility**: Sometimes there are compatibility issues with the versions of Python libraries or the specific setup of the environment that prevent certain code from running as expected.

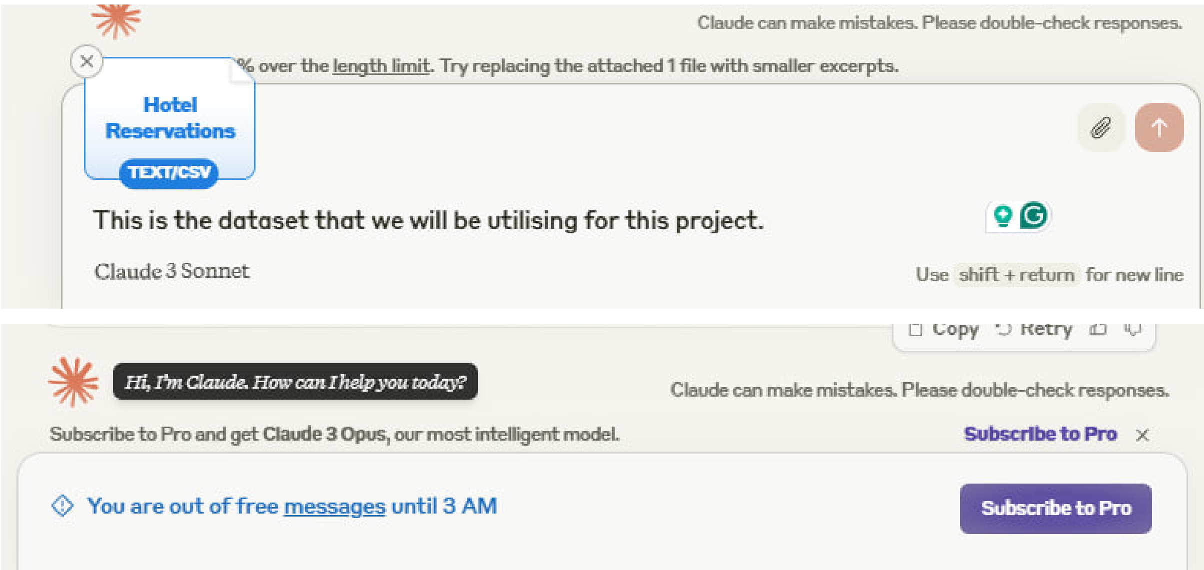
3. **Coding Errors**: Even though the logic might seem correct, there could be unseen errors in the code that are not handled, leading to exceptions that stop the code execution.

4. **Access and Permissions**: Certain operations might require access to external resources or specific permissions that are not available in the environment.

When these errors occur, I try to provide a conceptual explanation or a detailed code example that you can run in your own environment where you have full control over the execution context, libraries, and resource limits. This way, you can troubleshoot specific issues that arise, make necessary adjustments, and see the results directly.

If you have any specific questions about an error or need help debugging an issue you encountered while running the provided code, feel free to ask! I'm here to help guide you through solving these problems.

Appendix 3: Screenshot of Claude data upload and limited free prompts



Appendix 4: Full project directories, notebooks, codes can be found on https://github.com/AndreasL7/hotel_reservation/tree/main

Appendix 5: Data - how we utilise the metadata to do feature selection?

Booking_ID	unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer:
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not.

Appendix 6: Brief Comparison on Differences, Similarities and Strengths of each LLM (Secondary Research)

	GPT-4	Claude 3 Sonnet	Gemini Pro
Parameters	1.7 Trillion 120 layers	500 Billion	547 Billion
Cost	$100m	$100m	$700m
Undergraduate level knowledge	86.4%	79%	71.8%
Graduate level reasoning	35.7%	40.4%	-
Grade school math	92%	92.3%	86.5%
Math problem solving	52.9%	43.1%	32.6%
Multilingual math	74.5%	83.5%	63.5%

Code	67%	73%	67.7%
Reasoning over text	80.9	78.9	74.1
Mixed evaluations	83.1%	82.9%	75%
Knowledge Q&A	96.3%	93.2%	-
Common Knowledge	95.3%	89%	84.7%
Math & reasoning	56.8%	53.1%	47.9%
Document visual Q&A	88.4%	89.5%	88.1%
Math	49.9%	47.9%	45.2%
Science Diagrams	78.2%	88.7%	73.9%
Chart Q&A	78.5%	81.1%	74.1%
Number of top metrics	8/15	7/15	0/15

Source: https://encord.com/blog/claude-3-explained/