A REPORT

ON

MACHINE LEARNING PROJECTS AT STELLA STAYS

BY

ABHIGYAN GANDHI 2018A7PS0168U

AT

STELLA STAYS
DUBAI, UAE

A Practice School – II Station of

BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai
UAE

(AUGUST 2021 – JANUARY 2022)

A REPORT

ON

MACHINE LEARNING PROJECTS AT STELLA STAYS

BY

ABHIGYAN GANDHI 2018A7PS0168U CS

Prepared in Fulfillment of the
Practice School – II Course

AT

STELLA STAYS
DUBAI, UAE

A Practice School – II Station of

BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai
UAE

(AUGUST 2021 – JANUARY 2022)

BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai
UAE

Station: STELLA STAYS **Centre:** DUBAI

Duration: 02.08.2021 – 10.01.2022 **Date of Start**: 02.08.2021

Date of Submission: 10th January 2022

Title of the Project: MACHINE LEARNING PROJECTS AT STELLA STAYS

ID No. / Name of the student: 2018A7PS0168U / ABHIGYAN GANDHI

Discipline of Student: COMPUTER SCIENCE

Name(s) and Designation(s) of the Expert(s): ANANTA BAGHEL, PRDUCT LEAD
MUHANNAD SHUBITAH, SENIOR SOFTWARE DEVELOPER
ALI AL-SALIMY, PRODUCT SPECIALIST

Name of the PS Faculty: Dr. R SWARNALATHA

Key Words: Machine learning, feature engineering, pricing model, python, data science.

Project Area(s): Natural language processing, optical character recognition, sklearn models, tensorflow models.

Abstract (Max 200 words): This report gives details about my internship at Stella Stays, Dubai. I work with the tech and product team on individual machine learning projects which will be useful for the company. So far, I have done work on five different projects. Pricing model, sentiment analysis of reviews, OCR model, dynamic room allocation algorithm and automation of sending emails. I have built test models for the pricing model and the sentiment analysis model. For the OCR model I am using Google cloud services for their APIs, for the allocation algorithm I came up with the logic for it and also wrote a test script for it in python. I am currently working on the automation for sending emails.

R. Swarnalatha

Signature of Student **Signature of PS Faculty**

Date: 10th January 2022 **Date:** 10th January 2022

ACKNOWLEDGEMENTS

Firstly, I would like to express my heartfelt gratitude to Prof. Srinivas Madapusi, Director BPDC who has given us an opportunity to apply and understand our engineering concepts in a practical atmosphere.

I am grateful to Mr. Mohannad Zikra, CEO – Stella Stays, for assisting me by providing the required information about the organization.

I would like to thank Ms. Stephanie Ongaro, our PS Coordinator, for the smooth conduction of our internship and for guiding us throughout.

In my internship, I was mentored by three officials – Ananta Baghel, Muhannad Shubitah and Ali Al-Salimy. They all have helped me a lot, providing input and effort whenever needed. I am very grateful to have had the chance to work with them.

My sincere gratitude to Dr. R Swarnalatha, our PS Faculty, for providing me with all the assistance required for successful completion of this report.

I would also like to thank Dr. A Somasundaram, Dean, Practice School Division, for giving us this opportunity to work and apply our knowledge in the technical field and gain firsthand experience.

Abhigyan Gandhi
2018A7PS0168U

Table of Contents

List of Figures

Chapter-1 Introduction

1.1 About the Company

Figure 1.1 – Stella Stays logo

I am doing my PS-II internship at Stella Stays, Dubai. It is a hospitality brand which was founded in 2019. The CEO of the company is Mr. Mohannad Zikra, who is also one of the four co-founders. The company provides vacation rentals for short- and long-term stays. Currently they have properties in three locations, Dubai, Montreal, and Manama. The company was initially started in Montreal, Canada by the four co-founders. Then they decided to move to Dubai as the opportunity for growth here is massive. They started as a startup company and have been successful in establishing the brand as a leading high-quality enterprise in the hospitality sector. Their main office is also located here in Dubai where they have the greatest number of properties. They have various types of properties including villas, penthouses, and apartments available for renting. They also provide unfurnished apartments under the brand Ustella.

The company's motto is Redefining Hospitality, and they certainly have been doing that. Since the start, the company has been growing immensely. They have over 250 properties in Dubai, 13 properties in Montreal and 4 properties in Bahrain which they have just recently launched. The focus of the company is always on quality of the stay and the guests' experience. Whenever they acquire a property they do all the interior designing themselves and have a dedicated team for guest experience. The brand is becoming more and more popular every day and giving competition to other hospitality brands in the sector.

Working at the company has been wonderful for me so far. All the employees are very helpful and welcoming. They have a lot of teams including sales, design, marketing, product and tech, guest experience, and growth and development. It is always a fun environment in the office, and everyone is cheerful. The main goal right now is to expand and launch properties in different locations and regions. The most recent launch was in Manama, Bahrain which has been getting very good response from the guests.

1.2 Software and Platforms Used

I used the python programming language for all the projects as it is the most popular language for building machine learning models. It is very easy to understand and use as its syntax is very human readable and implementation is fairly simple. Machine learning models can easily be built using python because it has various open-source libraries which provide functions and implementations of a wide variety of machine learning techniques. I learned python on my own by doing projects in college, as we were taught other languages like Java and C in our college courses.

The development environment I use is the JetBrains PyCharm IDE. It is a great platform for developing projects because of its huge application features. Just using one platform you can create a virtual environment; access the command line and all the specific libraries you need for your project.

Figure 1.2 – PyCharm logo

For managing the datasets, I use Excel, which is a pretty standard software for data handling. A lot of feature engineering can be done using it, and from PyCharm you can directly export your results in the xlsx format which is very helpful.

For project management, updates and OKRs (Objectives and Key Results) the company uses ClickUp, which is easy to use and manage. You can create tasks, deadlines, goals and be updated with what is going on in the other teams for the project.

Figure 1.3 – Clickup logo

I have also used Google Cloud Services for various APIs, including the Vision API and the AutoML API for OCR ID verification and also the GMAIL API for sending automated emails.

Figure 1.4 – Google Cloud logo

Selenium framework was used for web scrapping and other automation processes.

Figure 1.5 – Selenium logo

Chapter-2 Projects

As I mentioned above, I have been assigned five different projects to work on individually. The pricing model, sentiment analysis, the OCR model, the dynamic room allocation algorithm and automation of sending emails. In this section, I will explain in detail about the research I have done and the progress I have made in each of the projects.

2.1 Pricing Model

So, the first project I got assigned was the pricing model. So currently for all the units the base price has been manually calculated and is changed by the sales team when the get an enquiry about a unit. The goal for this model is to predict the base prices for each unit based on certain factors. I started my research by looking at the prices of our competitors to see the trend of prices for units in different locations across Dubai. The most important thing was to decide the factors which will be used. By further research the factors I found were as follows:
Availability of units, our competitors' prices, season, festivals, discounts, location, amenities, internal expenses, target revenue, travel trends and restrictions, and length of stay. Out of these in my opinion the most important factors are our competitors' prices as we need to keep up with what they have to offer, in locations where there are no units available from the competitors we have a monopoly, so we can increase the prices. The next factor is season. The peak season for vacations in Dubai are the months from October to February, so we need to maintain the prices and hike them as per the season. We also need to meet a certain target revenue every month which is set by the sales team, so to make it easier for them, the base prices can be set according to the occupancy rate of the unit.

The datasets I was provided with included all the reservations which have been made till now, the daily base prices for each unit and the monthly target revenues. So first I had to feature engineering to get the dataset ready for putting it in a machine learning model. I made a test model for a single unit using a linear regression model. So firstly, I considered the daily base prices and used multipliers to get a final base price from the start of 2021 till October. I found out all the public holidays in UAE and created four categories, weekday, weekend, weekday plus holiday and weekend plus holiday, and set multipliers for them. Then I looked at the festivals which happen throughout the year and created multipliers for them. This way I got the final base prices. Other input factors I took into account were - a binary classifier for if the unit was on a night or not; the average price paid by the guests on a nightly basis for the previous bookings; and finally, I included the nationalities of the guests.

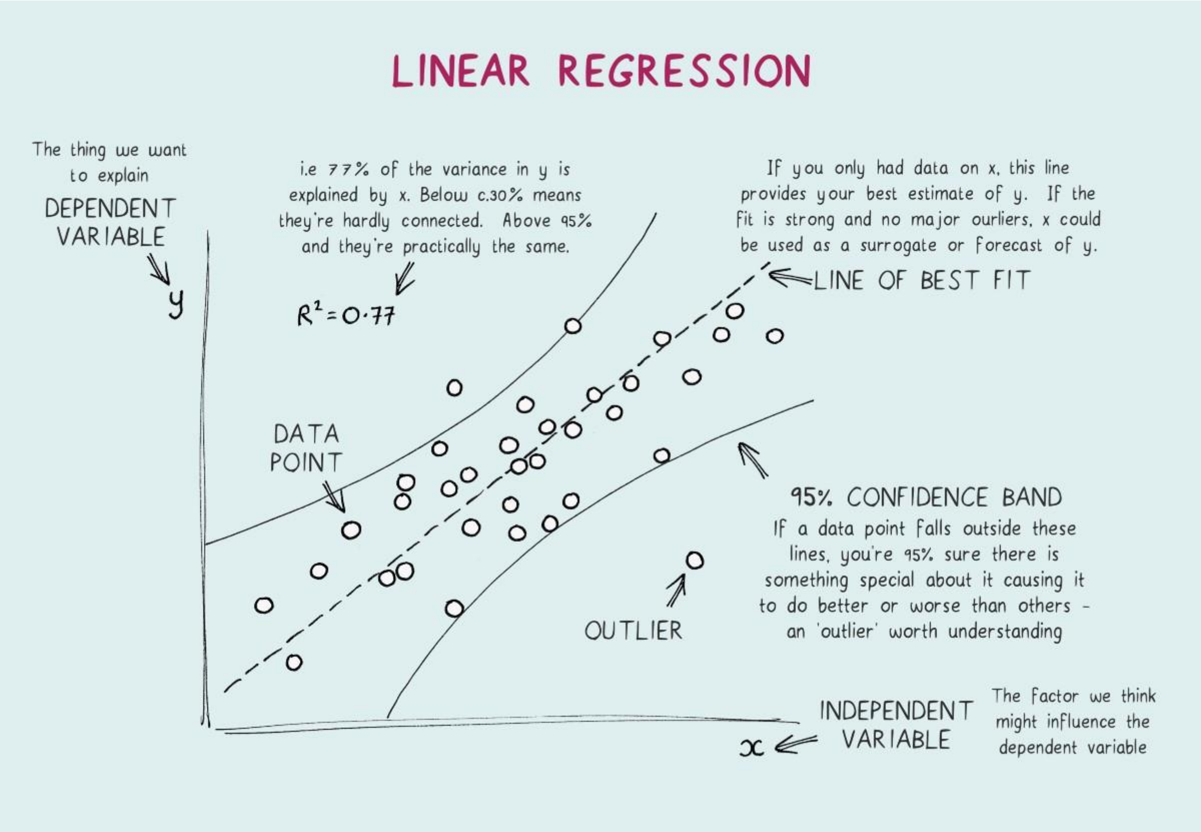

LINEAR REGRESSION
The thing we want to explain
DEPENDENT VARIABLE
y
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.
$R^2 = 0.77$
If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.
LINE OF BEST FIT
DATA POINT
95% CONFIDENCE BAND
If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others - an 'outlier' worth understanding
OUTLIER
INDEPENDENT VARIABLE
x
The factor we think might influence the dependent variable

Figure 2.1 – Linear regression

To implement this model, I used the sklearn python library which provides you with various machine learning techniques. I went ahead with the linear regression model as we do not have any complex variables and a linear equation can be formed easily. I split the dataset into test and train by using preprocessing function included in the sklearn library.

Further we will try to implement this model for all the properties and include more features like availability and location.

2.2 Sentiment Analysis of Reviews

After finishing with the test pricing model, the next project was to analyze the reviews we get and find out how the units are performing. So, I started my work by scraping the reviews from Airbnb by using the selenium library. This library is widely used for getting data from online platforms. It works by implementing a web driver of your choice, in my case I used the Google Chrome driver. Then you access a website, and using various functions get the data you need directly form the website and save it. Around 500 reviews were available on Airbnb which used as a test dataset. Now to train the model, I found a dataset on Kaggle, which had around 40,000 reviews of different hotels across Europe. First I had to assign each review a score and class.

The project area for this project is Natural Language Processing where you manipulate and analyze texts to obtain certain scores based on sentiments. I used the NLTK python library for this project as it has various applications. Using this library, you can process the data as per your requirements and score them. Firstly, I had to clean the data which means to get rid of less useful words, punctuation marks, non-alpha characters and stemming the words. This library has a list of stop words like a, an, the, we which are not useful while scoring reviews, so I removed them. Then I removed all the punctuations and digits and finally I stemmed all the words which means that words like eating, eaten and eats will change to eat.

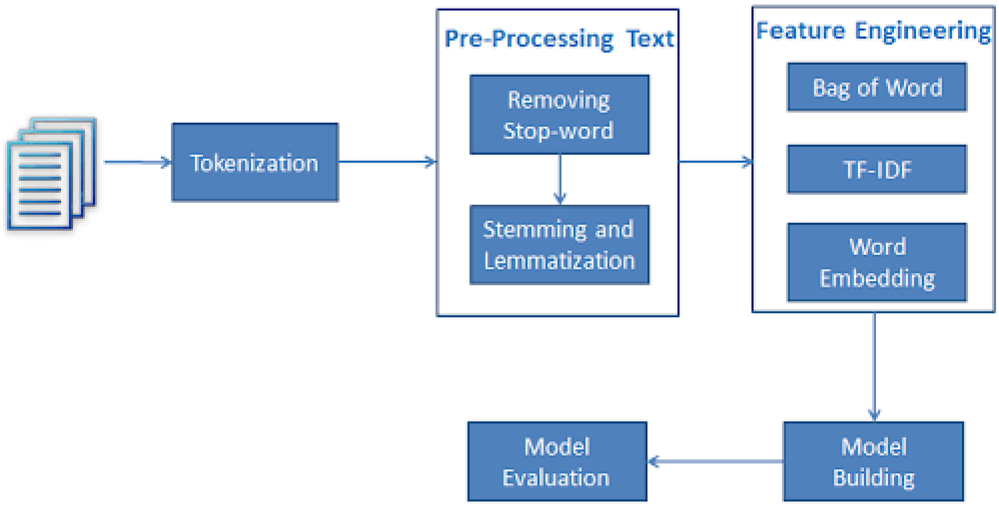


Figure 2.2 – NLP steps

After the reviews were ready I had to assign them scores which was done by the VADER sentiment intensity analyzer function available in the NLTK library. VADER stands for Valence Aware Dictionary for Sentiment. The VADER analyzer has a lexicon of words along with sentiment scores for each of them which were manually calculated. So, when you pass the review to the SentimentIntensityAnalyzer() function, it scores the review based on this lexicon and gives 4 different scores, positive, negative, neutral, and compound. After getting the scores, I assigned a class to each review based on the compound scores, so if the score was greater than 0.2 it was positive, if it was between -0.2 and 0.2 then it was neutral, and if it was less than -0.2 then it was negative. After assigning the scores to the train and test datasets, I decided to implement a tensorflow model to directly predict if a review is positive or negative. I used a keras sequential model which is basically a neural network for prediction. This model only takes numerical values as input so first I had to process the review to convert them into arrays of numbers which I did using the preprocessing functions available in the tensorflow library. I tokenized all the words and made padded sequences to use them as input for the model. The model had one hidden layer consisting of 8 nodes and an output layer using the sigmoid function as our output is a single variable. Currently there are some minor issues that need to be fixed for increasing the accuracy of the model.

After the testing is done, we deploy the model on our server. It will be able to scrape the reviews weekly from the websites and assign them scores.

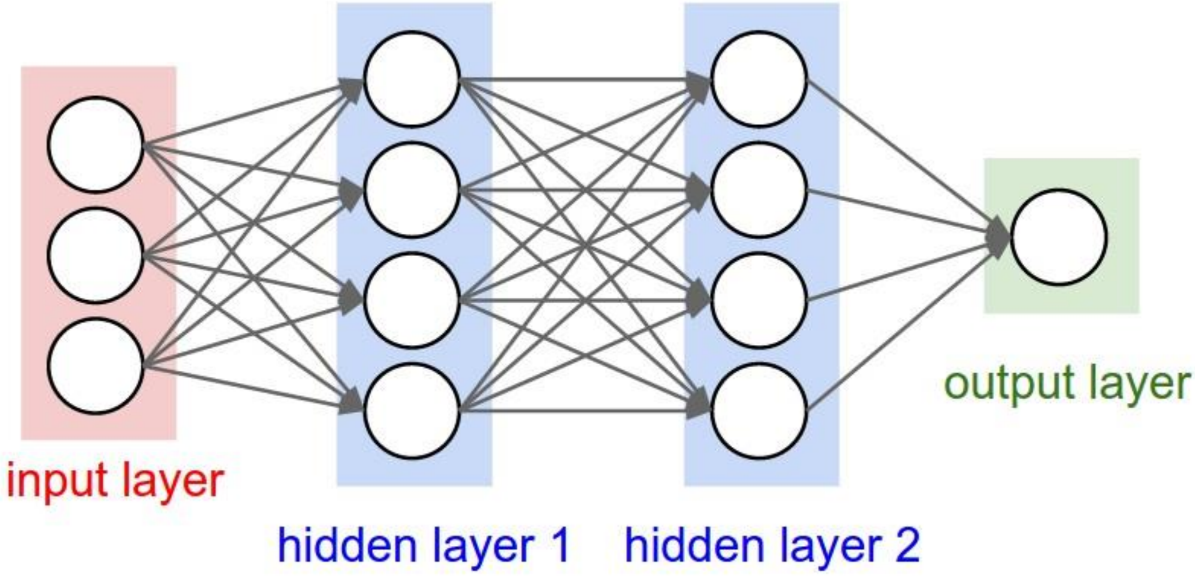


Figure 2.3 – Neural Network with eight nodes in two layers

2.3 OCR Model for ID Verification

The third project was for the automation for the verification of the IDs the guests send us. Currently the guest experience checks all the IDs manually to see if they are valid and legit. This task can be automated easily and will take a lot of load off of the GE team.

At first after doing some research and found the Pytesseract library which is an optical character recognition library, which basically means reading text from images. Using this library, we can input images and get the detailed text we need to verify the validity of the IDs. I tried to implement it, but the results were not that great. So, after doing some more research, I came across the Google Vision API, which offers pre-trained machine learning models for image annotations and manipulations. This API gave way better results and is easy to use, only downside is that it requires a monthly subscription to use through the Google Cloud Services.

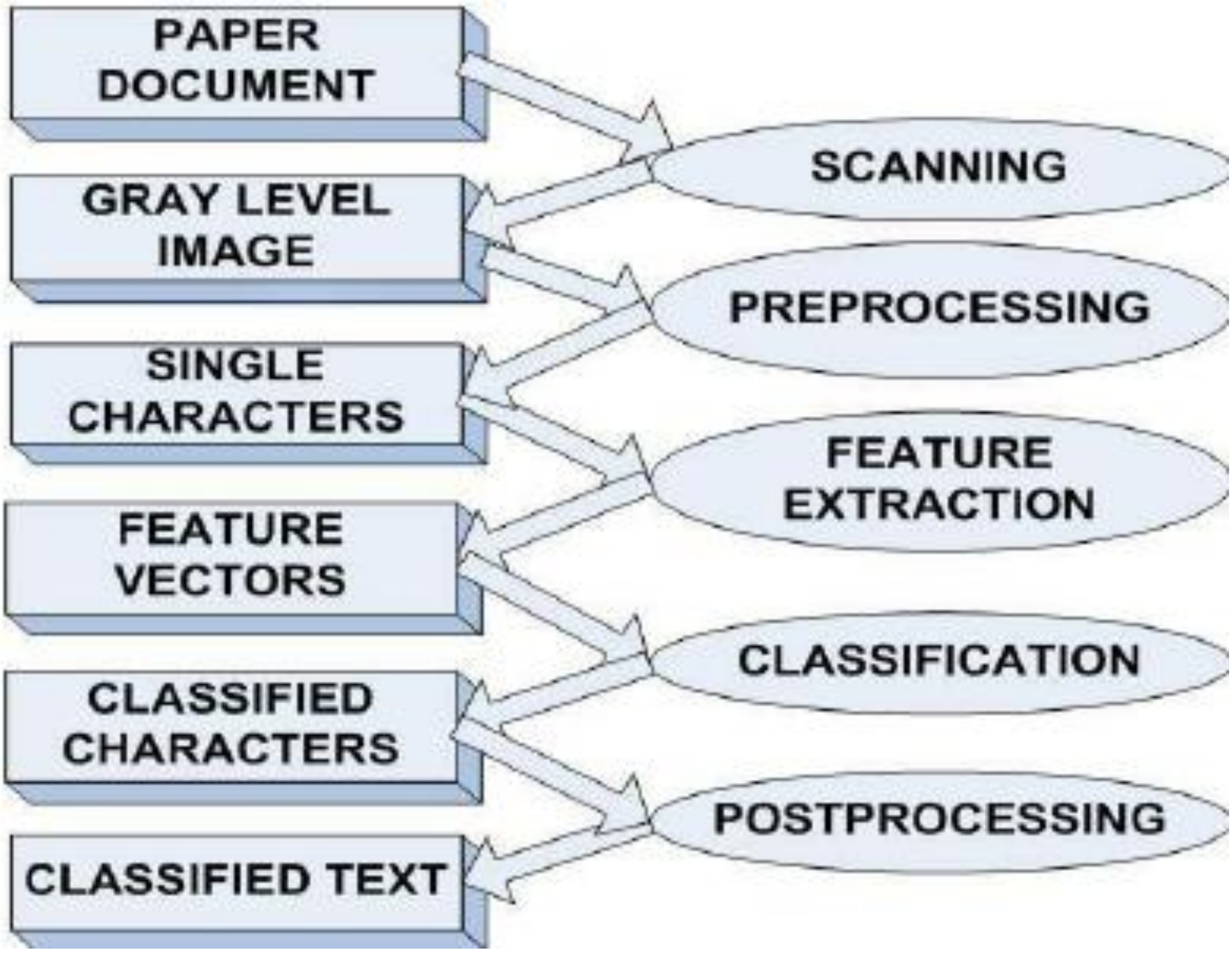


Figure 2.4 – OCR steps

Building an OCR model from ground up is very difficult as it requires high accuracy, and no mistakes can be made as the documents are official passports which cannot be interpreted wrongly.

So, for that reason we decided to look for third party OCR providers who provide SDKs for android and iOS. I did a lot of research and found some accurate and useful companies, the best out of them was Microblink who were providing what we needed. I had a meeting with them and discussed all the details and what will our requirements be. But we couldn't go ahead with them as they were not matching our feasibility.

After that I found out about the AutoML API provided by Google Cloud, it has pre-trained machine learning models for image annotation and manipulation. You can easily upload your dataset consisting of images of your IDs including passports and train the model and then get output based on your needs. In our case, we just needed details like name, age, passport number and address of the guest. We do not have enough data images to train the model for high accuracy, but it is a viable option to use.

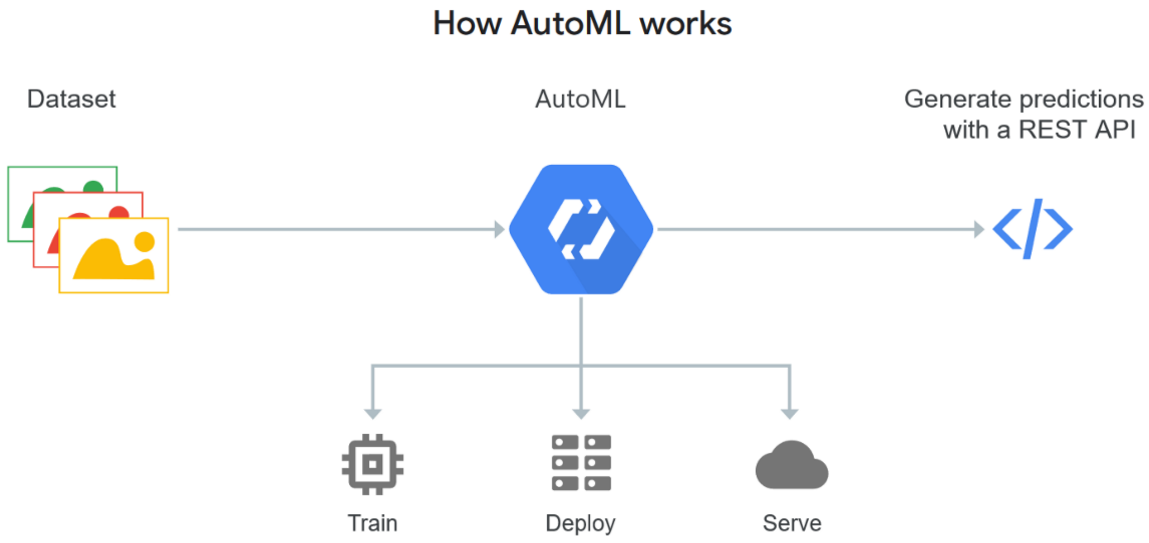


Figure 2.5 – AutoML API steps

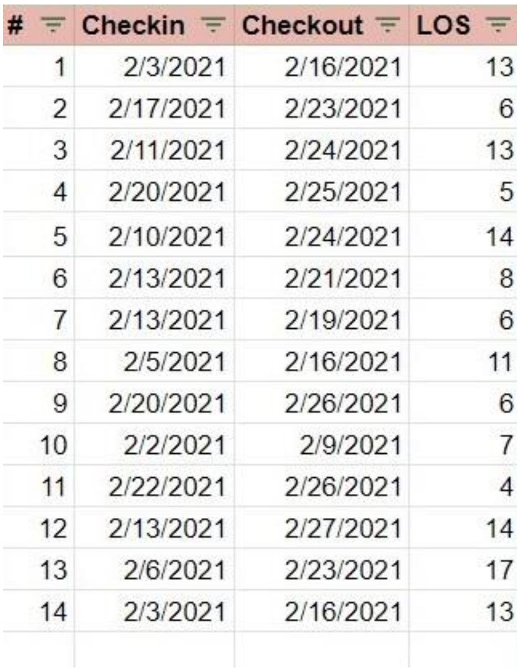
2.4 Dynamic Room Allocation Algorithm

The fourth project was for developing an algorithm for room allocation. With this project we wanted to create a hotel-like listing of our units. We are launching a new building in the Marina and a lot of rooms on the same floor are similar in terms of interiors, size, and amenities. So, like when you book a hotel room you do not get the room number you'll be staying in, you get the room number when you physically reach the hotel, so like that we wanted to create a single listing for the similar units and then assign the room number to the guest one or two days in advance from their check-in date.

I worked with Ali, our product specialist to come up with the logic for the algorithm. We did some research and came across the minimum slack concept. Now for the algorithm our priority was to minimize the gaps between two bookings to put them in the same room and also to maintain full occupancy rate at one unit rather than two units running at half occupancy rate as it helps in the maintenance and checking of units.

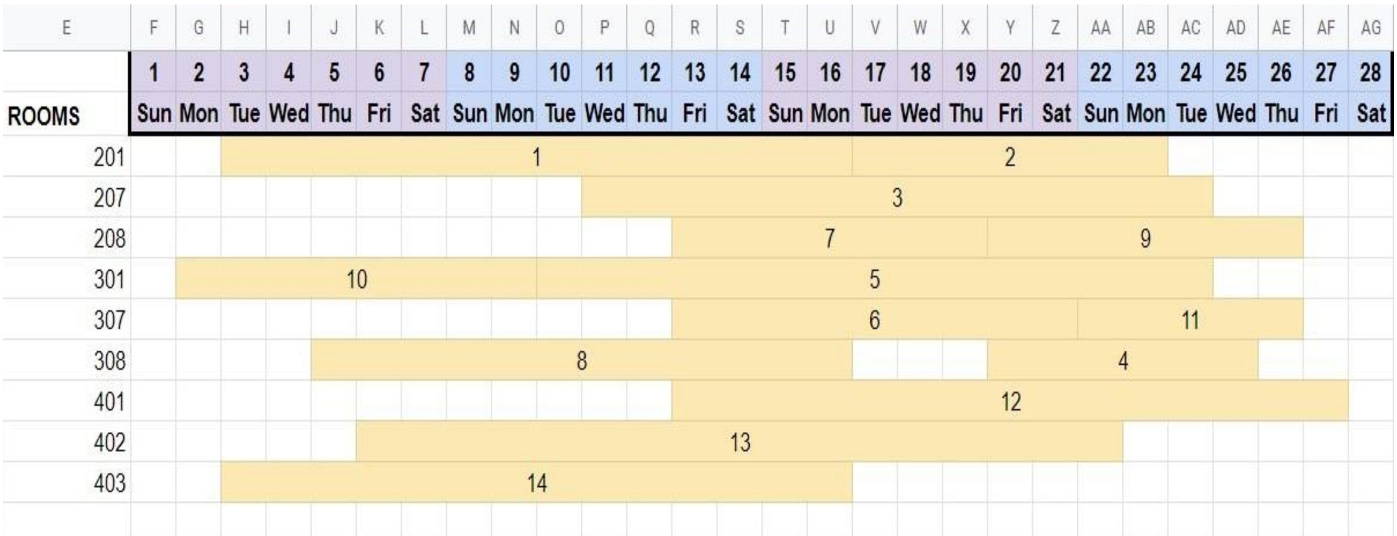
The logic which we developed first was that when a new booking comes in we will find the gap between that booking and all the previous bookings and with whichever booking the gap would be minimum the booking will be paired with it and inserted in the respective room. If two bookings have the same minimum slack then we took the top to bottom approach.

We took the following bookings coming in order and got this result shown in the image



#	Checkin	Checkout	LOS
1	2/3/2021	2/16/2021	13
2	2/17/2021	2/23/2021	6
3	2/11/2021	2/24/2021	13
4	2/20/2021	2/25/2021	5
5	2/10/2021	2/24/2021	14
6	2/13/2021	2/21/2021	8
7	2/13/2021	2/19/2021	6
8	2/5/2021	2/16/2021	11
9	2/20/2021	2/26/2021	6
10	2/2/2021	2/9/2021	7
11	2/22/2021	2/26/2021	4
12	2/13/2021	2/27/2021	14
13	2/6/2021	2/23/2021	17
14	2/3/2021	2/16/2021	13

Figure 2.6 – Input test bookings

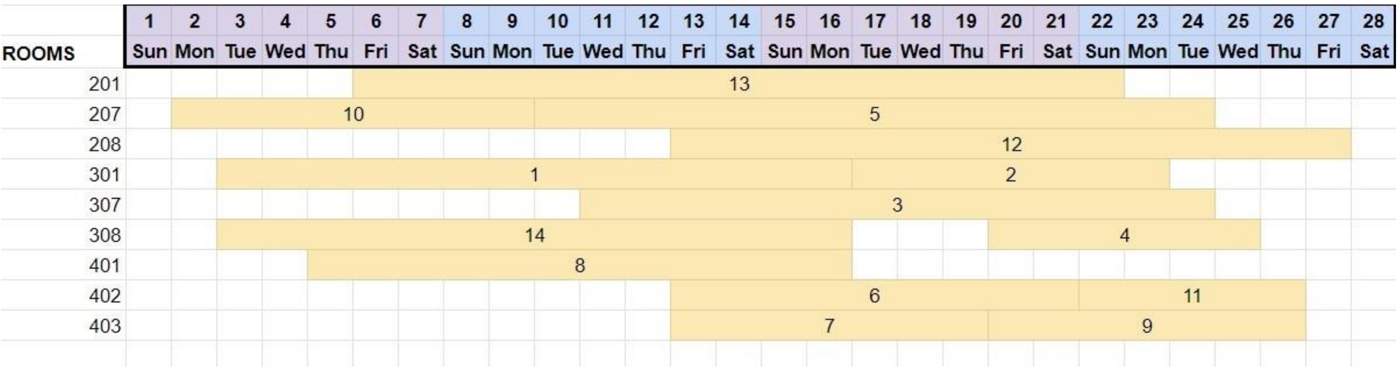


E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
ROOMS	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
201			▨	▨	▨	▨	▨	▨	1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	2	▨	▨	▨					
207											▨	▨	▨	▨	▨	▨	3	▨	▨	▨	▨	▨	▨	▨				
208													▨	▨	▨	7	▨	▨	▨	▨	▨	▨	9	▨	▨	▨		
301		▨	▨	▨	10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	5	▨	▨	▨	▨	▨	▨	▨				
307													▨	▨	▨	▨	6	▨	▨	▨	▨	▨	▨	11	▨	▨		
308					▨	▨	▨	▨	▨	8	▨	▨	▨	▨	▨	▨				▨	▨	4	▨	▨	▨			
401													▨	▨	▨	▨	▨	▨	▨	12	▨	▨	▨	▨	▨	▨	▨	
402						▨	▨	▨	▨	▨	▨	▨	▨	13	▨	▨	▨	▨	▨	▨	▨	▨						
403			▨	▨	▨	▨	▨	▨	14	▨	▨	▨	▨	▨	▨	▨												

Figure 2.7 – Output for minimum slack logic algorithm

After developing this logic, I worked some more on it and came up with another logic. This logic is based on clustering.

According to this logic whenever a new booking comes all the previous are taken out of the calendar and a cluster is formed using the overlapping bookings. By overlapping bookings, I mean all the bookings that have a common date on which the units are booked starting with the highest LOS booking. Then after the cluster is made it is put in the calendar in the rooms and then all the rest of the bookings are taken into account. We find the minimum slack of all the bookings and then start attaching them to the cluster in ascending order. At the end of this you get the following result on the same input data as above.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
ROOMS	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
201						▨	▨	▨	▨	▨	▨	▨	▨	13	▨	▨	▨	▨	▨	▨	▨	▨						
207		▨	▨	▨	10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	5	▨	▨	▨	▨	▨	▨	▨				
208													▨	▨	▨	▨	▨	▨	▨	12	▨	▨	▨	▨	▨	▨	▨	
301			▨	▨	▨	▨	▨	▨	1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	2	▨	▨	▨					
307											▨	▨	▨	▨	▨	▨	3	▨	▨	▨	▨	▨	▨	▨				
308			▨	▨	▨	▨	▨	▨	14	▨	▨	▨	▨	▨	▨	▨				▨	▨	4	▨	▨	▨			
401					▨	▨	▨	▨	▨	8	▨	▨	▨	▨	▨	▨												
402													▨	▨	▨	▨	6	▨	▨	▨	▨	▨	▨	11	▨	▨		
403													▨	▨	▨	7	▨	▨	▨	▨	▨	▨	9	▨	▨	▨		

Figure 2.8 – Output for clustering logic algorithm

2.5 Automation for Sending Emails

The last and latest project that I have been working on is for the automation of emails. This project is important as it will make this tedious task automated and easy for the GE team as they won't have to manually write an email for every check-in. So, for example if a guest checks in to a villa at The Palm Jumeirah, we need to fill in the DTCM (Department of Tourism and Commerce Marketing) form and guest details and send an email to Nakheel. This is a lengthy process and can be automated easily. The form filling also has to be automated but it very complex and we are figuring out the email part first.

So, for this I tried using the Selenium library which we have previously used for web scrapping reviews. This library can be used for multiple web automation needs. First I used the chrome web driver to access Gmail, but it didn't work as Google has halted automated processes from logging into human accounts.

So, I did some research and found the Gmail API available on Google Cloud Services. Using this API, we can easily send emails with attachments directly from the server.

This is the last project for my internship, and I have tested it out on the cloud console.

2.4 Other Tasks

While working in the office, sometimes we are required to help with various other tasks which are as follows:

So, for the release of the new version of the company’s website, they needed us to do product testing of the website which includes coming up with test cases and finding out bugs and errors in the websites’ flow. So, we helped with that and found some errors while testing every feature available on the website.

Another task I got was by the accounts department, they needed information about units in Dubai. Specifically, the number of nights a unit was booked in every month since January 2021. So, I used the reservations dataset which I was provided with for the pricing model and developed an algorithm to calculate the number of nights booked based on the check in and check out dates.

The last work I have left is to write the documentation for all the projects that I have worked on and hand over the files to the company which will be completed by the end of this week.

Chapter-3 Conclusion and Future Scope

3.1 Conclusion

In conclusion, this internship has been very useful since the beginning. I am starting to understand how a business organization works and have improved my skills a lot. I have been working on different projects and a lot of work has been done. In the pricing model, a test model for a single property has been created which is giving decent results. The sentiment analysis test model is complete and has been giving good results with decent accuracy. For the OCR project, we couldn’t find a third-party software feasible for us, so we went ahead with Google Cloud APIs, we just need more data images to train the model according to our needs and then get results. There has been a slight delay in the launch of our new building for which I was developing the room allocation algorithm, I made the test script, and it just needs to be tested for improving and error checking. And for the last project, it is pretty simple to implement and will be complete by the end of this week.

My experience working here has been amazing and productive. All our colleagues are very helpful and give the right guidance whenever required. The most important things which I have learnt are time management and how to be self-motivated. I still need to improve my technical and social skills further which I am certain will happen.

3.2 Future Scope

The goal for the pricing model was to automate the process of estimating the daily base prices of all properties based on various factors. Till now we have got a good measurement and understanding of how to go ahead. The test model is working fine, and it will be expanded to all the properties. For increasing the accuracy more factors will be included like the location, availability, and occupancy rate.

The sentiment analysis model is already in the test phase and when it will be accurate enough to function, it will be deployed on the server from where it will be able to get the reviews weekly and then classify and store them for analysis.

For the OCR model we are going ahead with the Google Cloud APIs as they also have a AutoML toolkit which has integrations with mobile applications. It just needs to be trained properly.

The dynamic room allocation algorithm is also in the testing phase, it needs to be given different inputs to check if it is working properly and make it error free.

The automation of emails is fairly easy now as we can use the Gmail API.

References

1. https://stackoverflow.com/
2. https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
3. https://www.siteminder.com/r/hotel-pricing/
4. https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python
5. https://monkeylearn.com/sentiment-analysis/
6. https://realpython.com/python-nltk-sentiment-analysis/
7. https://www.analyticsvidhya.com/blog/2021/06/sentiment-analysis-using-nltk-a-practical-approach/
8. https://towardsdatascience.com/a-complete-step-by-step-tutorial-on-sentiment-analysis-in-keras-and-tensorflow-ea420cc8913f
9. https://www.tensorflow.org/text/tutorials/text_classification_rnn
10. https://www.tensorflow.org/lite/examples/optical_character_recognition/overview
11. https://www.analyticsvidhya.com/blog/2020/05/build-your-own-ocr-google-tesseract-opencv/
12. Battiti, R., Brunato, M. and Battiti, F. (2021), "RoomTetris in room committing: why the role of minimum-length-of-stay requirements should be revisited", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 11, pp. 4017-4034. https://doi.org/10.1108/IJCHM-11-2020-1364
13. https://cloud.google.com/
14. https://www.selenium.dev/documentation/
15. https://www.edureka.co/community/2137/automate-gmail-login-process-using-selenium-webdriver-python
16. https://developers.google.com/gmail/api
17. https://developers.google.com/ml-kit/custom-models#automl_vision_edge
18. https://betterscientificsoftware.github.io/python-for-hpc/tutorials/python-pypi-packaging/