**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On

the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?

Hotel booking cancellations

* Why is this problem valuable to address?

Answer: Booking cancellations have a big impact on demand-management decisions in the hospitality industry. To associate with this problem, hotels implement rigid cancellation policies and overbooking strategies. However, this can also lead to negative influence on revenue and reputation.

* What is the current state (e.g. unsatisfied customers, lost revenue)?

Answer: Lost revenue.

* What is the desired state?

Answer: Increase revenue by reduce risks of cancellation from predicting which booking will be cancelled.

* Has this problem been addressed by other research projects? What were the outcomes?

Answer: Yes. This problem has been addressed by other research projects such as:

Antonio, Ana de Almeida and Luis Nunes, [Predicting hotel booking cancellations to decrease uncertainty and increase revenue](https://pdfs.semanticscholar.org/0f5f/3a506360b9be0a7ab52d77974695f1c48a4d.pdf) (2017)

# Industry/ domain

* What is the industry/ domain?

Answer: Hospitality/Hotel

* What is the current state of this industry?

Answer: This sector has been growing steadily over the past 5 years with the growth attributed to factors such as a growing ‘foodie’ culture, the popularity among many individuals to owning a hospitality business, population growth and an increase in consumer demand.

* What is the overall industry value-chain?

Answer:

Market Size: $12bn

Number of Businesses: 1,668

Industry Employment: 83,039

* What are the key concepts in the industry?

Answer: Hospitality.

* Is the project relevant to other industries?

Answer: No

# Stakeholders

* Who are the stakeholders?

Answer: Stakeholders are hotel managers and hotel revenue analysts.

* Why do they care about this problem?

Answer: Results from booking cancellation prediction can allow hotel managers to mitigate revenue loss from booking cancellations and to mitigate the risks associated with overbooking.

* What are the stakeholders’ expectations?

Answer: Stakeholders expect to have model that can predict booking cancellations and what elements affect it.

# Business question

* What is the main business question that needs to be answered?

Answer: How can we increase hotel revenue by predicting booking cancellations?

* What is the business value of answering this question? (quantify value and make necessary assumptions)

Answer:

* What is the required accuracy? What are the implications of false positives or false negatives?

Answer: Model should achieve a prediction accuracy above 0.8 and roc\_curve (AUC) also above 0.8.

# Data question

* What is the data question that needs to be answered?

Answer: Based on previous bookings, which machine learning model could be used to predict hotel booking cancellations?

* What is the data required to answer the question?

Answer: Bookings database of hotels

# Data

* Where was the data sourced?

Answer: The data was obtained from “Nuno Antonio, Ana de Almeida and Luis Nunes, [Hotel booking demand datasets](https://www.sciencedirect.com/science/article/pii/S2352340918315191) (2019)”.

* What is the volume and attributes of the data?

Answer: 119,390 samples and 32 features.

* How reliable is the data?

Answer: The data is reliable.

* What is the quality of the raw data?

Answer: Raw data contains data of four hotels which was concatenated together in this project. This created a large number of missing values. Some features are not useful in predicting the target.

* How was this data generated?

Answer: The data was collected directly from the hotel’s PMS databases using Microsoft SQL Server.

* Is this data available on an ongoing basis?

Answer: Yes. More data can be generated from PMS when needed.

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?

Answer: Dealing with missing values, outliners, feature engineering, data preprocessing

* What are the highlights of the Exploratory Data Analysis (EDA)?

Answer: \_High cancellation rate in 3 years.

\_No deposit bookings have the highest cancellations.

\_Bookings made a few days before the arrival date are rarely cancelled, whereas bookings made over one year in advance are cancelled very often.

* Is the pipeline reusable? (for example, to process future data?)

Answer: Yes.

* What are the intermediary data structures used (if any)?

Answer: None

## Modelling

* What are the main features used?

Answer: 'hotel', 'is\_canceled', 'lead\_time', 'arrival\_date\_year',

'arrival\_date\_month', 'arrival\_date\_week\_number',

'arrival\_date\_day\_of\_month', 'stays\_in\_weekend\_nights',

'stays\_in\_week\_nights', 'adults', 'children', 'babies', 'meal','country', 'market\_segment', 'distribution\_channel',

'is\_repeated\_guest', 'previous\_cancellations',

'previous\_bookings\_not\_canceled', 'reserved\_room\_type',

'assigned\_room\_type', 'booking\_changes', 'deposit\_type', 'agent','company', 'days\_in\_waiting\_list', 'customer\_type', 'adr',

'required\_car\_parking\_spaces', 'total\_of\_special\_requests',

'reservation\_status', 'reservation\_status\_date'

* Did you find any interesting interactions between features?

Answer: Yes.

* Is there a subset of features that would get a significant portion of your final performance? Which features?

Answer: Yes. The most important features are lead\_time, adr, deposit\_type

* How did you select features?

Answer: Features are selected by two ways: manually by the researcher based on business knowledge and experience, RandomForest’s feature\_importance\_

* What feature engineering techniques used?

Answer: Pandas get dummies for categorical features and StandardScaler for numerical features.

* What are the models used?

Answer: RandomForestClassifier, DecisionTree, XGBoost, LogisticRegression.

* How long does it take to train your model?

Answer: Less than 5 minutes.

* What are the tools used? (cloud platform, for example)

Answer: Jupiter Notebook, libraries: Numpy, Pandas, Seaborn, Matplotlib, Sklearn.

* What are the model performance metrics?

Answer: Accuracy, Precision, Recall, roc\_curve (AUC)

* Which model was selected?

Answer: RandomForestClassifier was selected as it has the best accuracy at 89%.

## Outcomes

* What are the main findings and conclusions of the data science process?

Answer: The model can predict target with over stakeholder’s expectation. Some insights from EDA are useful in understanding booking cancellations and what factors have influence on it.

## Implementation

* What are the considerations for implementing the model in production?

Answer:the booking cancellation prediction model should not be implemented by itself. In truth, if deployed independently of the hotel other systems, it is unlikely that it would present any valid results in terms of revenue management. Today’s speed and complexity imposed on a hotels reservations department is such that advantages of using the model could not be clear if tasks related to the model inputs and outputs had to be done manually. For that reason, the model should be integrated on the hotel CRS. This will enable the CRS to have more accurate net demand forecasts and consequently, present better overall forecasts. (Nuno Antonio, Ana de Almeida and Luis Nunes, [Predicting hotel booking cancellations to decrease uncertainty and increase revenue](https://pdfs.semanticscholar.org/0f5f/3a506360b9be0a7ab52d77974695f1c48a4d.pdf) (2017)).

# Data answer

* Was the data question answered satisfactorily?

Answer: Yes. I was able to build a machine learning model that can predict hotel booking cancellations.

* What is the confidence level in the data answer?

Answer: The data answer has high confidence as it surpasses the stakeholder’s expectation.

# Business answer

* Was the business question answered satisfactorily?

Answer: Yes. By having a model to predict booking cancellations, hotel managers can reduce risks of booking cancellations by offering hotel services to customers who are likely will cancel. Establishing better overbooking strategies and improving cancellation policies. This has potential to create more sales, since less rigid cancellation policies can generate more bookings.

* What is the confidence level in the business answer?

Answer: The business answer has a high confidence level and potential in increasing hotel’s revenue.

# Response to stakeholders

* What are the overall message and recommendations to the stakeholders?

Answer: The model is able to predict booking cancellations for hotels. However, results may change with different hotel types, stars, locations, special events and natural disasters. Hotel managers need to be careful when dealing with the False Positive rate which is quite low at 12,4% in total bookings. This means 12,4% of total bookings were predicted cancelled but actually did not cancel. Inapproriate handlings of this number could lead to loss of revenue and bad reputation for hotel.

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

Answer: By using RandomForestClassifier model, I was able to solve the business and data question. The model accuracy surpassed the stakeholder’s expectation at 0.89 which means it’s good enough to be implemented into real life business.

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

Answer: The data and code used in the project can be found at

<https://github.com/DSIA-Education/2019_12_10_DS_PT_SYD_Projects/tree/master/henry/capstone>

* What are the resources used in the project? (libraries, algorithms, etc)

Answer: Numpy, Pandas, Sklearn, Seaborn, Matplotlib.