



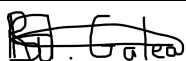
## DISSERTATION SUPERVISION LOGBOOK

<b>Institute</b>	Institute of Information & Communication Technology
<b>Programme</b>	B.Sc. Business Analytics (Hons.)
<b>Dissertation Title</b>	<b>Forecasting sales for aesthetic products using Machine Learning</b>
<b>Supervisor</b>	<b>Alan Gatt</b>
<b>Student</b>	<b>Rushayeal Galea Massa</b>
<b>Student ID No</b>	<b>364201L</b>

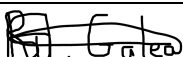
### Note

- I. It is the **student's responsibility** to ensure that this logbook is correctly documented and maintained, and that Supervisor recommendations and signatures are acquired after each and every meeting.
- II. This logbook is to be submitted together with the dissertation.
- III. The institute reserves the right **to not accept** the student's dissertation for evaluation if this logbook is **not filled in correctly** and **duly signed** by the student and supervisor as indicated.



<b>Meeting Number : 1</b>		<b>Date of meeting : 13/10/2022</b>
<b>Issues discussed at the meeting (<i>to be filled in by Student</i>)</b>  SOI draft finished, to include more details and information.		
<b>Supervisor recommendations (<i>to be filled in by Supervisor</i>)</b>  Revise SOI as per comments found on Word Document. Overall, include more research papers, define objectives better, and focus the research questions.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
27/10/2022		A. Gatt



<b>Meeting Number : 2</b>		<b>Date of meeting : 27/10/2022</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>SOI submitted. To start working on subheading and structuring the Literature Review in a better way. GitHub structure was explained and to be completed by next meeting. For the next tutor session, start working and exploring the data in Python.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<ul style="list-style-type: none"><li>• Start working on literature review (pick headings from SOI and start fleshing them out). Identify headings that need to be written about.</li><li>• Start analysing data, by loading it in Python, and check for missing data, any transformations needed, and graphs. Create in Notebook – Data Exploration.</li><li>• On GIT create structure:<ul style="list-style-type: none"><li>○ Doc/ (dissertation)</li><li>○ Src/ (source code)</li><li>○ Papers/ (papers)</li><li>○ Data/ (data file)</li></ul></li></ul>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
3/11/2022		A. Gatt



<b>Meeting Number : 3</b>	<b>Date of meeting : 3/11/2022</b>
<b>Issues discussed at the meeting (<i>to be filled in by Student</i>)</b>  <p>SOI feedback was provided by the mentor. The literature review subheadings listed in the SOI were broken down into more subtitles before the meeting. These were reviewed together with the mentor and made some changes to find them useful when writing the literature review.</p> <p>The data set was discussed, and the mentor allocated some small tasks needed to start doing the data cleaning, which need to be completed by the next few meetings.</p>	
<b>Supervisor recommendations (<i>to be filled in by Supervisor</i>)</b>  <u>Subtitles:</u>  2.1 Machine Learning: 2.1.1 What is ML 2.1.2 Difference between supervised and unsupervised  2.2 Forecasting sales or demand: 2.2.1 Why forecasting is used 2.2.2 Time Series modelling  2.3 Importance of sales forecasting: 2.3.1 Reaching customer demand 2.3.2 Seasonality 2.3.3 Shelf life 2.3.4 (Aesthetic documentation)  2.4 Covid-19 affecting sales:	



2.4.1 Worldwide lockdown

2.4.2 Covid period in UK

To do:

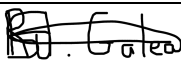
Add new column to ex. 1 nov 2015

Delete Sales rep column

Change Location to actual location

Rename products + add a category

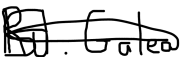
Rename customer & company

Date of Next Meeting	Student Signature	Supervisor Signature
24/11/2022		A. Gatt

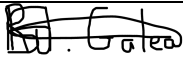


<b>Meeting Number : 4</b>		<b>Date of meeting : 24/11/2022</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>When writing the code, a problem was encountered. The data needed to be anonymized but the loop to generate the unique names had an error. This was revised together with the mentor and the companies and clients' names were changed to general unique names. Further improvements to the script were discussed.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Discussed problems with data cleaning.</p> <p>Revised script to generate unique names, and assign a label to remove identity of companies and clients.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
2/12/2022		A. Gatt



<b>Meeting Number : 5</b>		<b>Date of meeting : 2/12/2022</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>  Data Cleaning source code was revised and finalized. Few columns which are not going to be used need to be removed, afterwards the clean dataset needs to be exported as a csv. A suggested structure for the source notebooks was given by the mentor. The next recommendations are to explore the data using PowerBi visuals and make note of what can be used when exploring with Python. It was also noted to start working on the literature review.		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>  Remove columns from dataset which will not be used e.g. Data Entry Date, Month-Year, Year Quarter  Create a structure where each notebook performs a specific task: <ul style="list-style-type: none"><li>- Notebooks<ul style="list-style-type: none"><li>o Data Cleaning</li><li>o Data Exploration</li><li>o Regression Model</li></ul></li></ul> Start working on Data Exploration; export csv from data cleaning, load in PowerBI and make notes on what can be interesting.  Start literature review.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
15/12/2022		A. Gatt



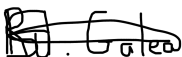
<b>Meeting Number : 6</b>		<b>Date of meeting : 15/12/2022</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>More work needs to be completed in the PowerBI file. It was noted that visuals can be broken down to separate the dominant values over the smaller ones. Two different visuals can be created to help in understanding the data more easily.</p> <p>It was also recommended to take different approaches on different slides to list down different observations and to be able to make an analysis at the end.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Continue working on PowerBI. Go through each page, and if Perfilho is dominant try to break the visual into two visuals (one for Perfilho and the other for the rest).</p> <p>In the initial slides you can take an overall approach (2d comparison e.g. year vs qty).</p> <p>In the subsequent slides try to gather more information using a 3d approach (e.g. year vs qty vs category).</p> <p>Write down observations, and maybe include tables (e.g. Category vs Year showing values of total quantity).</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
12/01/20203		A. Gatt





<b>Meeting Number : 7</b>		<b>Date of meeting : 12/01/2023</b>
<b>Issues discussed at the meeting (<i>to be filled in by Student</i>)</b>  It was recommended to finish up the PowerBI visuals and continue working on the literature review.  Research needs to be done on how to predict time series data when working with categorical values such as our data set.		
<b>Supervisor recommendations (<i>to be filled in by Supervisor</i>)</b>  PowerBI work is still not finished.  Some more work was done on the dataset.  Continue working on the literature review.  Research how to predict time series data when having categorical values.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
26/01/20203		A. Gatt



<b>Meeting Number : 8</b>		<b>Date of meeting : 26/01/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Literature review needs to be refined by adding automatic referencing in Word. More research needs to be found on which machine learning algorithms are used to predict demand and supply and write two paragraphs on them. The paragraphs need to include at least 4 references.</p> <p>PowerBI visuals need to be plotting against the date to attempt to find out seasonality in the data. For the charts to be easily interpreted, it was recommended to separate the products first. Map visuals needs to be created to show frequencies of orders and amount of sales while also having filters of products and dates.</p> <p>Research needs to be completed on how to predict time series when having categorical values.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Literature Review – Find existing research on which machine learning algorithms are used to predict demand and supply. Write a paragraph or two on this having at least 3 or 4 references.</p> <p>Change referencing to automatic in Word.</p> <p>PowerBI – Attempt to find seasonality by plotting against date. Separate products so that charts can be easily interpreted. Map to show frequency of orders, amount of sales, filter by product, filter by date.</p> <p>Research how to predict time series data when having categorical values.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
13/02/2023		A. Gatt



<b>Meeting Number : 9</b>		<b>Date of meeting : 13/02/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Referencing was updated to automatic, while research about machine learning algorithms still needs to be completed.</p> <p>PowerBI visuals need to be analysed to list important observation in them by using the smart narrative tool. The important observations need to be listed in a document.</p> <p>Some research needs to be carried out on whether LabelEncoding or OneHotEncoding is the best and most used technique.</p> <p>Some columns need to be removed from the dataset while also making some other small changes.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Literature Review – Find existing research on which machine learning algorithms are used to predict demand and supply. Write a paragraph or two on this having at least 3 or 4 references.</p> <p>Referencing changed to automatic mode.</p> <p>PowerBI – Pages are finished, with observations extracted. Try to use smart narrative on every page to extract anomalies. Try changing filters / clicking on elements to see how observations change. Prepare a document with all the “important” observations.</p> <p>Research how to predict time series data when having categorical value; research existing papers to check what is done such as LabelEncoding and OneHotEncoding.</p> <p>Dataset: Quarter remove Q. Remove column Cust.Name, Comp.Name, Area Code and Code.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
27/02/2023		A. Gatt

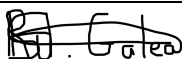


<b>Meeting Number : 10</b>		<b>Date of meeting : 27/02/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Literature review needs to be finished.</p> <p>PowerBI observations which were extracted need to be further explained and given more context.</p> <p>Time-series tutorials needs to be practiced and understood.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Literature Review still in progress.</p> <p>PowerBI – Observations are extracted. It is suggested to go through all the observations and give them more context. For example:</p> <p><i>Sum of Qty trended up, resulting in a 24,077.78% increase between Sunday, November 1, 2015 and Tuesday, March 1, 2022.</i> In this case you can mention how the quantity started and ended in actual figures rather than percentages.</p> <p><i>Sum of Qty started trending up on Monday, February 1, 2021, rising by 613.44% (1871) in 393 days.</i> In this case you can then observe what happened after the increase. Did it stay the same, decrease?</p> <p>Same for rest of observations...</p> <p>Practice time-series regression on tutorials.</p> <p>Dataset should be cleaned.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
09/03/2023		A. Gatt



<b>Meeting Number : 11</b>		<b>Date of meeting : 09/03/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>  Literature review finished, to be reviewed and given feedback by the lecturer.  PowerBI Observations still need to be explained further and given more context.  Tutorials should first be practiced on the dataset of the tutorials, then tried on own dataset.  Research: 'demand forecasting', 'predicting supply chain', 'forecast demand over time'.  Data cleaning file needs to be updated.		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>  Final preparations of data cleaning.  Revise Power BI statements.  Perform found tutorial as is, then try to apply on your data.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
16/03/2023		A. Gatt



<b>Meeting Number : 12</b>		<b>Date of meeting : 16/03/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>  Literature Review still needs to be reviewed by lecturer.  PowerBI Observations still need to be explained further and given more context.  Tutorials having regression on multiple columns should be tried out.		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>  Still need to review literature review.  Continue on tutorials that have regression on multiple columns rather than a single one.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
23/03/2023		A. Gatt



<b>Meeting Number : 13</b>		<b>Date of meeting : 23/03/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Literature Review needs to be reviewed and updated according to the lecturer's feedback.</p> <p>PowerBI Observations still need to be explained further and given more context.</p> <p>More tutorials need to be tried out on my own dataset and evaluate the results.</p> <p>A paragraph containing relevant research needs to be completed.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Literature Review feedback needs to be reviewed and updated accordingly.</p> <p>PowerBI Observations still need to be explained further and given more context.</p> <p>More tutorials should be tried out on own dataset.</p> <p>Relevant research paragraph needs to be completed.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
30/03/2023		A. Gatt



<b>Meeting Number : 14</b>		<b>Date of meeting : 30/03/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Literature Review was reviewed during the meeting to explain the comments, this needs to be reviewed and updated according to the lecturer's comments.</p> <p>PowerBI Observations are finished.</p> <p>A paragraph containing research related to supply/demand needs to be completed.</p> <p>Current tutorials need to be finalized, then repeat the experiment using different variations. After a different algorithm needs to be tried out.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Went through some of the comment of the literature review.</p> <p>Finish paragraphs found from research papers related to supply / demand.</p> <p>Finalise current experiment. Then repeat experiment considering different variations of data:</p> <ul style="list-style-type: none"><li>• Starting with all columns</li><li>• Removing variety of columns</li><li>• Then repeat previous 2 on a particular category only (in this case category column must not be used)</li></ul> <p>Repeat experiments with a different algorithm</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
20/04/2023		A. Gatt





<b>Meeting Number : 15</b>		<b>Date of meeting : 20/04/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Discussed results of experiment.</p> <p>The mentor gave a step by step instructions on the next few steps to be taken while trying experiments.</p> <p>More experiments using different algorithms should be tested out.</p> <p>Some changes to be made in the pipeline.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Improve current experiment by using RandomizedSearchCV (<a href="https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html">https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html</a>) / GridSearchCV (<a href="https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html">https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html</a>)</p> <p>Duplicate experiment, filter out products by category, and repeat experiment for each category.</p> <p>After this has been done, use a different algorithm than RandomForest (at least 2; e.g., XGBoost, and Neural Networks).</p> <p>To check literature review.</p> <p>Pipeline -&gt; Data exploration -&gt; Test Data, Training Data -&gt; Train Model (connect Test Data to it) -&gt; Evaluate Model (loop back to Train Model) -&gt; Results</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
27/04/2023		A. Gatt



<b>Meeting Number : 16</b>		<b>Date of meeting : 27/04/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>		
<p>Prepare draft of the results based on the RMSE obtained from each experiment for Rainforest, XGBoost and Neural Networks. Compare results both when using and not using filters. Write down the best parameters for each experiment.</p> <p>Start drafting methodology covering, Data acquisition, data description, data cleaning, data exploration, data preparation for machine learning, training of models.</p> <p>Try another experiment by splitting the data by random and split data to find a particular section.</p>		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>		
<p>Draft Results parts</p> <p>First compare results of all categories (without filtering) and write down results for each different algorithm.</p> <p>Repeat the same thing but with filtering.</p> <p>For each model listed above, write down the chosen parameters.</p> <p>Methodology:</p> <p>Data acquisition, data description, data cleaning, data exploration, data preparation for machine learning, training of models.</p> <p>Experiments:</p> <p>Split data by random, split data to find a particular section.</p>		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
4/05/2023		A. Gatt



<b>Meeting Number : 17</b>		<b>Date of meeting : 4/05/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>  Worked and fixed the experiment which splits the data randomly.  It was suggested to try a different experiment of the prototype which will be able to test and train on selected period of time. First try to choose a covid year as test data, then try summer or winter period, and then a whole year as the test data.  Also, revise literature review while continuing to work on methodology.		
<b>Supervisor recommendations (to be filled in by Supervisor)</b>  Fixed experiment that splits data randomly.  Next task is to try different experiments by selecting a period of time. For instance, choose the covid year as test data, a summer or winter period as test data, a whole year as test data.  Continue working on methodology, and revise literature review.		
<b>Date of Next Meeting</b>	<b>Student Signature</b>	<b>Supervisor Signature</b>
11/05/2023		



<b>Meeting Number : 18</b>	<b>Date of meeting : 11/05/2023</b>
<b>Issues discussed at the meeting (to be filled in by Student)</b>  The literature review needs to be revised and split up into subheading as below. Some additional paragraphs need to be added to introduce the subheadings.  Methodology needs to be continued.  Final experiment to be tried: Pick a year, pick covid year, pick summer / winter of last year.  All of above should be run on two different algorithms and also filtering should be included for optimal results.	
<b>Supervisor recommendations (to be filled in by Supervisor)</b> Revise literature review:  Literature Review:  (Introduction - sales forecasting using machine learning algorithms related to cosmetic products) 2.1 Sales Forecasting (1.1) (introduction) 2.1.1 Meeting customer demand (1.1.1) 2.1.2 Aesthetics products supply and demand (1.1.4) 2.1.2.1 seasonality (1.1.2) 2.1.2.2 shelf life (1.1.3) 2.1.3 Extraordinary events affecting supply and demand (2.4) (introduction) 2.1.3.1 Covid-19 (2.4.1) (combine under this header - worldwide lockdown, Covid in the UK) 2.1.3.2 Brexit (2.4.2) 2.1.3.3 change in government (2.4.3) 2.2 Machine Learning (2.1) (introduction) 2.2.1 Introduction to machine learning (2.1.1) 2.2.2 ML Approaches (2.1.2)  2.3 Forecasting supply and demand (2.2) 2.3.1 Time Series Modelling (2.2.2) (remove from sub-heading include with 2.2.2 time series modelling with categorical values) 2.3.2 forecasting models (2.2.4) 2.3.3 related studies (2.2.5)  Methodology – to continue  Final experiments to run:  Pick a year, pick covid year, pick summer / winter of last year	



All of above should be run on two different algorithms / filtering should be included for optimal results

Date of Next Meeting	Student Signature	Supervisor Signature
25/05/2023		A. Gatt

Meeting Number : 19	Date of meeting : 25/05/2023
<p><b>Issues discussed at the meeting (to be filled in by Student)</b></p> <p>The results document needs to be restructured, remove the random splits and add the last months of the data instead.</p> <p>In each splitting section, the RMSEs need to be written down for each algorithm and each category.</p> <p>For each also do a table with actual, predicted and difference and put in results (covid, last months) and the rest in appendix (summer, winter, whole year).</p> <p>A table needs to be created at the end of the results to list the lowest RMSE and the algorithm name for each category and splitting period.</p> <p>In the appendix also put all the results obtained.</p>	
<p><b>Supervisor recommendations (to be filled in by Supervisor)</b></p> <p>Re-structure results part. Have these sections:</p> <p>12 months of 2019</p> <p>summer 2019</p> <p>winter 2019</p> <p>covid year</p> <p>Last 4 months of 2022</p> <p>In each section, do a table for the RMSEs for each category for each algorithm.</p> <p>For covid year, and 2022 include a table with actual / predicted – for the rest put in appendix.</p> <p>In the appendix add algorithm parameters for all.</p> <p>Do a table at the end with all the categories vs sections, with the lowest RMSE and algorithm type.</p>	

Date of Next Meeting	Student Signature	Supervisor Signature
		A. Gatt