A Regression Problem:
The Best Time to Purchase an Airline Ticket

Weerada Sattayawuthipong
220345604
Fabrizio Smeraldi
MSc Big Data Science

***Abstract*—Airline ticket prices are extremely fluctuating, which makes it difficult for customers to find the best period to purchase it. To answer this question, web scraping scripts were developed to gather airline ticket flight data for routes from London to five cities in Asia, data preprocessing, outlier removal, and regression models were applied to find the period providing the minimum price of the airline tickets. For flights going to Bangkok, Seoul, and Singapore, an early purchase strategy could save more money while for Hong Kong and Tokyo flights, the customers better wait until a few days before departure to buy the ticket.**

Keywords— airfare, airline, ticket, regression, linear, isotonic

I. INTRODUCTION

The airline industry is highly competitive. The airline's objective is to expand the revenue generated while the customer intends to reduce the cost of traveling. The most common tool airline companies use is a dynamic price that pairs between supply and demand profiles (Groves and Gini 2011). The quantity of available cheap-fare tickets is reduced if the actual sales are better than expected and the available quantity of cheap-fare tickets is increased when the actual sales do not meet the expectation (Wen and Yeh 2017). The ticket price for the same route could be varied within one day. For example, a flight from Los Angeles to Boston could be changed up to approximately 7 times per day (Etzioni et al. 2003). These are just factors from the airline's strategy. The other factors affecting the demand for ticket prices could come from special or unpredictable events happening in specific locations such as concerts, conferences, disasters, and pandemics (Abdella et al. 2021). Recent research discovered the effect of COVID-19 on the airline industry which reduced more than 50 percent of the number of worldwide passengers and flight traveling distances due to travel limitations, border closures, and distancing approaches (Wozny 2022). These uncertain factors could lead the airlines to lose their revenue and the customers to face the risk of buying airline tickets too early before the actual departure date. Several research has been proposed to avoid this unpleasant situation for airline customers. Most research focused on the route within Europe or America, however, this research intended to investigate more on the routes from London to particular capitals in Asia.

The primary objectives of this project are to study the airline ticket price patterns for specific routes and the best time to purchase the lowest ticket price possible. The models presented in this paper described the relationship between days before departure and the proportion of cost savings using a regression model. The result comparison between each route was also addressed. The subsequent sections of the paper are structured as follows: background research is presented in the next section, followed by methodology and data collection. Finally, the result and discussion together with a conclusion and possible future research are addressed.

II. BACKGROUND

Discovering the best time to buy an airline ticket is a challenging problem due to the airline's price strategy and uncertain conditions. By exploring related work, previous research can be categorized into classification and regression problems. The research solving a classification problem provided a model suggestion to perform or delay the ticket purchase while the research solving a regression problem provided the optimal period of buying an airline ticket.

The founding research for the classification problem was the work of Etzioni et al. (2003). In this research, the model was built based on ticket price observation of 21 days in advance for 41-day periods of non-stop, round-trip flights from Los Angeles to Boston and Seattle to Washington DC. Some interesting ticket price behavior was found. For instance, flights departing around holidays appear to fluctuate more than non-holiday flights, and ticket prices were usually increased two weeks before departure dates and were at peak on the departure dates. The researchers developed a data mining algorithm named "Hamlet" to generate a model suggesting to customers whether to buy or wait for a specific airline ticket compared to the predictive minimum airline ticket price. This algorithm used results combination from Q-learning, Ripper (Etzioni et al. 2003), and Time series analysis. By comparing the performance of Hamlet with other methods, such as optimal, handcraft, or Ripper model, the proposed model showed a desirable outcome compared to other methods resulting in 4.4% savings of ticket prices on average over 4,500 simulated passengers. Another research with a similar wait-and-buy approach was contributed by Groves and Gini (2011). The study extended the period before flight departure to 60 days in advance. Strong cyclic patterns were found in the dataset. Round-trip passengers who departed on Monday and returned on Friday tended to travel for business purposes while passengers who took different days for round-trip flights tended to travel for holiday. Business flights were more insensitive to price, enabling airlines to raise ticket prices early without significantly reducing demand. The model in this study was constructed based on a Partial Least Squares (PLS) Regression that allows users to select the number of features when training the model and it is robust to highly collinear or irrelevant features. Model performance was validated based on a comparison of price saving from the model itself to the immediate purchase approach and airline ticket price prediction from Bing Travel. The result revealed that the model could reduce the average expense of purchasing airline tickets in the 60 days period ahead of the departure date. Later, the optimal purchasing time for airline tickets using a decision support service was created by Xu and Cao (2017). The system was developed based on both historical and real-time airline ticket price data. Moving Average, Classification, and Regression Trees were used to create a multi-step airline ticket price purchase strategy called Dynamic Potential Days with Lower Price (DPLP) (Xu and Cao 2017). This work compared results between DPLP with other purchase strategies, such as Potential Days with Lower Price (PLP), and Optimal Stopping Rules (OSR) (Xu and Cao 2017), based on loss and gain from purchasing airline tickets starting 40 days prior to departure date. The proposed system delivered a satisfying performance. Additionally, this work stated that the probability of buying lower airline ticket prices decreased as the days before departing decreased.

Outstanding regression problem research finding the optimal time to buy airline tickets was achieved by Domínguez-Menchero et al. (2014). This research collected airline ticket prices 30 days prior to the departure date for a period of 60 days for 4 specific routes. The researchers applied several models, such as linear, isotonic, and cubic regression, to identify the relationship between cost-saving percentage and the number of days before departure. The isotonic regression provided an obvious range of time for delaying airline ticket purchases and how much customers could save from purchasing airline tickets before a certain period. The optimal time for purchasing an airline ticket without any major financial penalty in this work appeared to be at least 18 days before the departure date.

In this paper, the same approach as in Domínguez-Menchero et al. (2014) work was applied to finding the optimal period of purchasing airline tickets. The isotonic regression is the only model that provides a clear range of time customers could purchase the minimum airline tickets (Domínguez-Menchero et al. 2014), hence this project focused on fitting flight data using isotonic regression. To observe the ticket price trend during the given period, linear regression was applied as the initial method. Details of the model fitting will be explained in the following section.

III. Methodology

All methodologies used in this project were implemented using Python programming language. For model fitting, a machine learning package: scikit-learn (Pedregosa et al. 2011) was used. Other specific Python packages will be addressed in the subsections.

A. *Web Scraping*

Web Scraping could be referred to as an automatic procedure extracting data from websites by transforming unstructured web data into structured data which could be stored and evaluated in any data storage such as a spreadsheet or database (Khder 2021). This technique was developed to exclude unwanted data and obtain only the desired data (NR et al. 2023). In this project, the web scraping process was accomplished using a browser automation tool: Selenium with Python packages: Beautiful Soup and regular expression. Selenium controlled a web driver object, Google Chrome, to open a given URL and obtained its contents (NR et al. 2023). Beautiful Soup built a parse tree for parsing HTML and XML files to extract specific data (NR et al. 2023) while regular expression determined a desired pattern in a string and extracted for any matches of that string (Khder 2021).

Using the above concept, the process started by scraping airline ticket flight data from https://www.kayak.co.uk/ once a day at different times. A list of URLs for individual routes and departure dates was generated to fetch these URLs to Selenium. A sample URL was in this format: https://www.kayak.co.uk/flights/LON-BKK/2023-04-22?sort=bestflight_a. The ticket flight data consists of destination, departure date, search date, ticket price, airline name, and number of flight connections. Route and departure date data were taken from the URLs. Ticket prices, airline names, and number of flight connections were extracted from the HTML in Table I. using Beautiful Soup and applied regular expression in Table II. These data were later stored in CSV files as columns for further pre-processing and fitting of the model. In case the scraping failed which resulted in the file size being less than 1020 bytes, the program would start scraping again until the size reached the condition. The web scraping flow was displayed in Fig. 1.

TABLE I. HTML FOR TICKET PRICES, AIRLINE NAMES, AND NUMBER OF FLIGHT CONNECTIONS.

Data	HTML
Ticket price	<div class="f8F1-price-text">£456</div>
Airline name	<div class="J0g6-operator-text">Thai Airways</div>
Number of flight connections	<span class="JWEO-stops-text">direct</span>

TABLE II. REGULAR EXPRESSION FOR TICKET PRICES, AIRLINE NAMES, AND NUMBER OF FLIGHT CONNECTIONS.

Data	Regular Expression
Ticket price	re.search('(\<.+?\>)(.+?)(\<.+?\>)', str(price_lst[i])).group(2)
Airline name	re.search('(\<.+?\>)(.+?)(\<.+?\>)', str(airline_list[i])).group(2)
Number of flight connections	re.search('(\<.+?\>)(.+?)(\<.+?\>)', str(direct_list[i])).group(2)

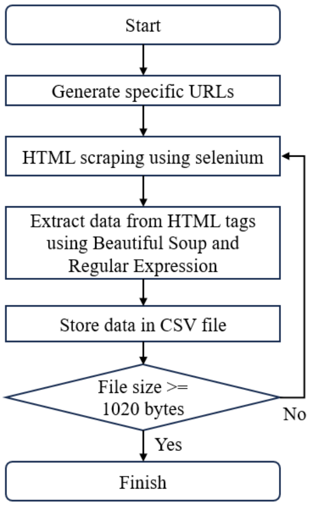
Start
Generate specific URLs
HTML scraping using selenium
Extract data from HTML tags using Beautiful Soup and Regular Expression
Store data in CSV file
File size >= 1020 bytes
No
Yes
Finish

Fig. 1. Web scraping flow.

B. Saving Rates

A concept to estimate how much cost could be saved if purchasing a ticket before the departure date was inspired by the work of Domínguez-Menchero and team (2014) by calculating the saving percentage of the price on the search date compared to the price on the departure date. Let p_0 be the lowest price on the departure date and p_l be the price on the search date, where l = -66, -65, -64, …, -1 represented the number of days before the departure date. For example, l = -66 denoted 66 days before the departure date. The saving rate was represented as:

$$\text{saving rate} = \left(\frac{p_0 - p_l}{p_0}\right) * 100 \tag{1}$$

C. Data Preprocessing and Outlier Removal

To simplify the process of calculating the saving rate, the scraped CSV files were grouped by departure route into folders d0 and dx using the Pandas Python package. Folder d0 contained files that were scraped on the departure date representing price p_0 from (1) while folder dx contained files that were scraped on the day before the departure date representing price p_l from (1). New CSV files were generated for each departure route consisting of 11 columns including the original column from the web scraping step and 5 additional columns containing preprocessing depart date, number of days before departure date, day of the week, price on departure date (p_0) and saving rate. Firstly, data cleaning was applied to remove commas and the special character "Â£" from the ticket price column. The preprocessing depart date and search date columns were formatted as 'YYYY-MM-DD'. Then, the number of days before the departure date was a subtraction between the search date and departure date. Lastly, the saving rate was calculated using the formula in (1) from the previous section.

Rapid increases or decreases in ticket prices could be identified as outliers which were defined as "A data object that deviates significantly from the normal objects as if it were generated by a different mechanism" (Han et al. 2023). Referring to Kwak and Kim's work (2017), outlier removal should be applied to the data set to prevent the circumstance of overestimating or underestimating the model and it could be detected from the distance between a data point and the center of all data points using median and quartile range which were less sensitive to outliers compared to mean and standard deviation. Statistically, outliers are defined as "the value above 1.5 times the interquartile range (IQR) of the third quartiles (Q3) or the value below 1.5 times the IQR of the first quartiles (Q1)" (Han et al. 2023).

Referring to Han et al. (2023), a median is a value at a central position of an ordered dataset in case the total amount of data in the dataset is odd and it is the average of two most middle values if the total amount of data in the dataset is even. The calculation of the median was shown in (2).

$$\text{Median} = \begin{cases} \mathrm{X_{n+1}} & \text{if N is odd; } \mathrm{N = 2n + 1} \\ \frac{1}{2}(\mathrm{X_n + X_{n+1}}) & \text{if N is even; } \mathrm{N = 2n} \end{cases} \tag{2}$$

The definition of the IQR from Han et al. (2023) is a distance between Q3 and Q1 displayed as in (3) indicating how the middle half of the data spread. Q3 is the 75th percentile taking the lowest 75% of the sorted data while Q1 is the 25th percentile cutting off the lowest 25% of the sorted data.

$$\text{IQR} = \text{Q3} – \text{Q1} \tag{3}$$

D. Linear Regression

A regression model is a method used to predict and determine the causal relationship between independent and dependent variables. Linear regression is the simplest model estimating a dependent variable 'y' with a single independent variable 'x' (Maulud and Abdulazeez 2020). The true relationship between these variables could be displayed as in (4) where β_0 is a constant value known as an intercept, β_1 denotes slope, and ε represents the noise.

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \varepsilon \tag{4}$$

As in the article of Maulud and Abdulazeez (2020), a least square method is a solution for the linear regression problem. Its concept is to find the best-fit line for a true y using a prediction value $\hat{y} = b_0 + b_1 x$ that reduces the cumulative squared distance between $\hat{y}$ and y as possible. It has been proved that b_0, b_1 could be found by solving (5) and (6), where $\bar{y}$ is a mean value calculated from all $y_i \in y$ and $\bar{x}$ is a mean value calculated from all $x_i \in x$.

$$b_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

E. Isotonic Regression and Simplification Process

Isotonic regression is "an example of nonparametric regression where the number of parameters grows linearly with the number of data points" (Kotłowski et al. 2016). Referring to the definition from Barlow et al. (1978), giving a finite set X = $\{x_1, \ldots, x_k\}$ with the order $x_1 \prec x_2 \prec \cdots \prec x_k$, a real valued function f on X is isotonic if x_i, x_{i+1} and $x_i \prec x_{i+1}$ imply $f(x_i) \leq f(x_{i+1})$. Suppose g is a function on X and *w* defined as positive function on X. An isotonic function $\hat{g}$ on X is an isotonic regression of g with weight *w* regard to the simple ordering $x_1 \prec x_2 \prec \cdots \prec x_k$ if it minimizes

$$\sum_{x_i \in X} [g(x_i) - f(x_i)]^2 w(x_i)$$

Generally, isotonic regression fits a non-decreasing function (Kotłowski et al. 2016), however it can fit a non-increasing function f on X = $\{x_1, \ldots, x_k\}$ with the simple order $x_1 \prec x_2 \prec \cdots \prec x_k$ in case of x_i, x_{i+1} and $x_i \prec x_{i+1}$ imply $f(x_i) \geq f(x_{i+1})$ using Spearman's rank correlation coefficient (r_s) (Pedregosa et al. 2011). Referring to Gautheir (2001), r_s is a "nonparametric technique evaluating the degree of linear correlation between two independent variables which operates on the rank of data". The concept is to separately rank each variable from smallest to largest value and record the change between ranks of each data pair. If there is a correlation in the data, the sum square of the difference between ranks will be low. A calculation of r_s shows in (7). Where d_i denotes the change between ranks for each data pair (x_i, y_i) and *n* is the number of data pairs.

$$r_s = \frac{1 - 6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{7}$$

The predictions' function $\hat{g}$ of the isotonic regression form a piecewise linear function (Pedregosa et al. 2011) which is used in simplification process. Similar to Domínguez-Menchero et al. (2014) work, the simplified process is to find the turning points from the isotonic function $\hat{g}$ and smooth several piecewise linear functions between these turning points into a single linear function using (5), (6) in section E. For example, for Bangkok route, a turning point p1: (-30,6.4916751), p2: (-22,4.22890137), p3: (-18,4.22890137), and p4: (-1,1.10957961) were observed. Therefore, three simplified isotonic functions were created between pairs of points p1-p2, p2-p3, and p3-p4 visualizes in Fig 2. The same approach also applied to the other routes.

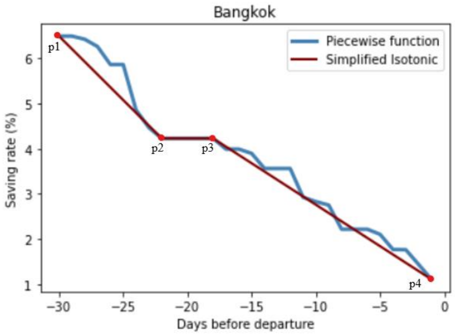

Fig. 2. Piecewise function from isotonic regression and simplified isotonic function for Bangkok route 30 days prior flight data.

F. Goodness-of-fit Metrics

Measurement metrics are tools that compare the prediction values $\hat{y}$ and the actual values y. They assess the model quality by measuring a deviation between $\hat{y}$ and y related to the concepts of distance and similarity (Botchkarev 2019). In this paper, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) were used to assess the sense of good fit.

Mathematical expression, stated in the work of Chicco et al. (2021), for MSE, MAE, and R^2 for data set with N samples are presented in (8) - (10) where $\hat{y}_i$ is a prediction values from a model, y_i is an actual value, and $\bar{y}$ is a mean value calculated from all y_i. While MSE finds the average error using a squared error between $\hat{y}$ and y, MAE employs absolute error to prevent offsetting of positive and negative errors (Botchkarev 2019). R^2 is a result of one minus ratio of the variance explained by the linear model to the total variance (Plevris et al. 2022). Referring to the explanation from Chicco et al. (2021), values of MSE and MAE could be between 0 to +∞ where 0 indicated the model fitting with no error while the values of R^2 is between $-\infty$ to $+1$ in which +1 implied perfect prediction. MSE and MAE could be used for comparing which model has better perform than the others, however, they are difficult to interpret their value alone compared to R^2 value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{8}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2} \tag{10}$$

IV. Dataset

The flight data of specific airlines with directed flights departing from London and arriving at Bangkok, Hong Kong, Tokyo, Seoul, and Singapore were scraped for model fitting. The airline ticket price data were gathered 30 days and 66 days before the departure date.

For ticket flight data collected 30 days earlier to the departure date, the airline ticket data were collected for 39 different departure dates starting from 22 April to 30 May 2023 using the Web Scraping technique mentioned in section III. This data was collected 30 days prior to the departure date until the day of the departure. For example, for a departure date of 22 April 2023, the airline ticket flight data were gathered every day from 23 March until 22 April 2023. Due to a scraping issue encountered on 28 April 2023, the airline ticket prices for the Seoul route were excluded from the dataset. The same approach was used to gather ticket flight data 66 days prior to the departure date, except that this dataset was gathered only for 7 different departure dates starting from 24 to 30 May 2023. The number of samples and airlines' names after outlier removal for each destination are presented in Table III.

TABLE III. THE NUMBER OF SAMPLES AND AIRLINES' NAMES AFTER OUTLIER REMOVAL

Data sets	**Routes**	**Number of samples**	**Airlines**
Data gathered 30 days before departure	Bangkok	1844	EVA Air, Thai Airways
	Hong Kong	3146	British Airways, Cathay Pacific
	Tokyo	806	ANA, British Airways
	Seoul	2138	Asiana Airlines, Korea Air, Virgin Atlantic, KLM
	Singapore	3103	British Airways, Singap ore Airlines, Qantas Air ways
Data gathered 66 days before departure	Bangkok	760	EVA Air, Thai Airways
	Hong Kong	1052	ANA, British Airways
	Tokyo	333	British Airways, Cathay Pacific
	Seoul	829	Asiana Airlines, Korean Air, Virgin Atlantic, KLM
	Singapore	1585	British Airways, Singap ore Airlines, Qantas Air ways

V. Results & Discussion

Equations estimated a linear regression for ticket flight data collected 30 and 66 days before departure of each destination is displayed in Table IV, where $\hat{y}$ represents a predicted saving rate from the model and *x* represents number of days before departure. The results of the linear regression for each destination are visualized in Fig. 3 and Fig. 4. The saving rate was decreasing when the day of departure came closer for flights to Bangkok, Seoul, and Singapore which is opposite to the flight to Hong Kong and Tokyo that purchasing ticket closer to departure date could save more money.

TABLE IV. EQUATIONS RESULTED FROM LINEAR REGRESSION

Datasets	Routes	Equations
Data gathered 30 days before departure	Bangkok	$\hat{y} = 1.08 - 0.18x$
	Hong Kong	$\hat{y} = 1.27 + 0.15x$
	Tokyo	$\hat{y} = 1.77 + 0.03x$
	Seoul	$\hat{y} = 2.05 - 0.13x$
	Singapore	$\hat{y} = 7.30 - 0.40x$
Data gathered 66 days before departure	Bangkok	$\hat{y} = 2.55 - 0.10x$
	Hong Kong	$\hat{y} = -7.27 + 0.04x$
	Tokyo	$\hat{y} = 4.09 + 0.05x$
	Seoul	$\hat{y} = 2.05 - 0.16x$
	Singapore	$\hat{y} = 8.24 - 0.20x$

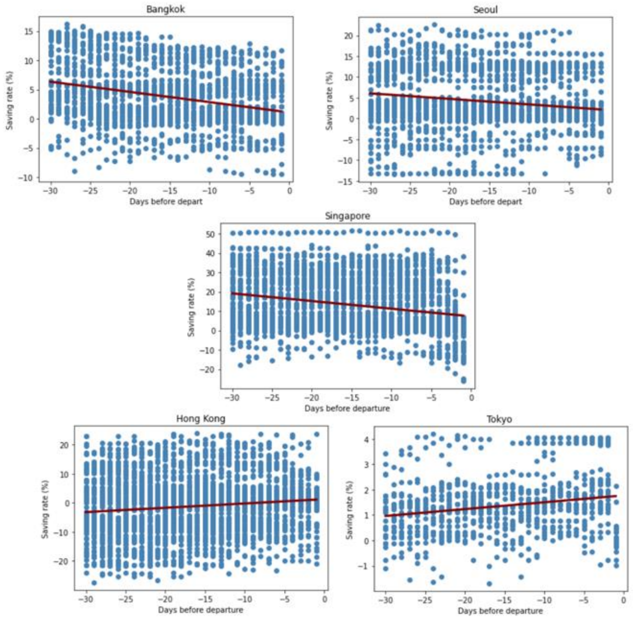

Fig. 3. Linear Regression for ticket flight data gathered 30 days prior to the departure date.

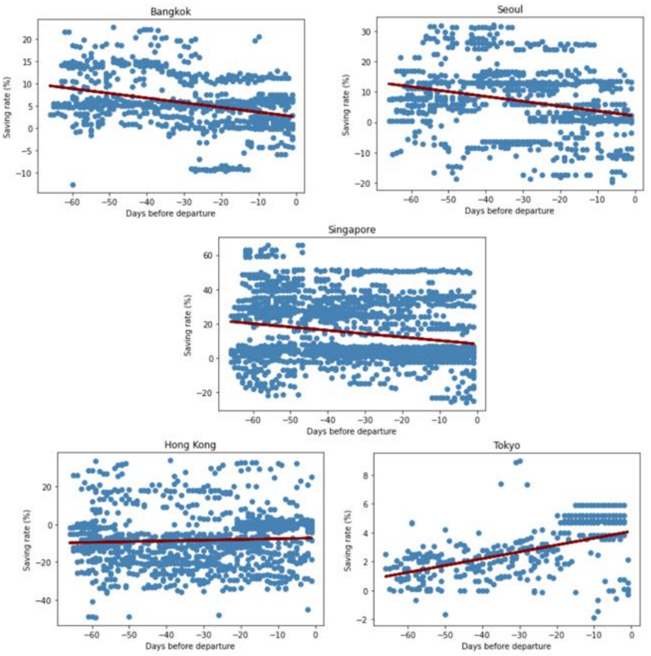

Fig. 4. Linear Regression for ticket flight data gathered 66 days prior to the departure date.

The equations yielded from simplified isotonic regression for each route are displayed in Table V, where $\hat{y}$ represents a predicted saving rate from the model and x represents number of days before departure. The figures demonstrating isotonic regression and simplified isotonic regression for ticket flight data gathered 30 and 66 days before departure date of each destination are shown in Fig. 5 - Fig. 8. Referring to the result of the simplified isotonic regression, it is obvious that the earlier airline ticket purchase could save more money for Bangkok, Seoul, and Singapore routes while it is opposite for the other two routes.

For ticket flight data gathered 30 days ahead of the departure date, the predicted saving rate for Bangkok maximized on 30 days before departure at 6.41% and it stayed at 4.23% during 18 to 22 days prior to the departure date before it went down to approximately 1% on the day prior to the departure date. The highest saving rates for flight to Seoul and Singapore were forecasted on 30 days before departure date at 5.55% and 17.19 %, and rapid decrease trends of these routes were noticed on the 10 and 6 days before the departure date until they reach the lowest saving rate at below 1% for both destinations. For Hong Kong and Tokyo flights, the saving rate increased gradually from 30 days before the departure date and remained the same after 8 and 10 days ahead of the departure date at 0.69% and 1.67%, respectively.

For ticket flight data gathered 66 days before the departure date, the best time to purchase an airline ticket to get the maximum saving rate for Bangkok and Seoul routes was 66 to 36 days before the departure date which resulted in 8.33% and 10.39%. For the flight going to Singapore, the ideal period for buying a ticket was 66 to 47 days prior to departure date giving 18.83% saving rate whereas for the trip to Tokyo, purchasing a ticket after 18 days before departure date resulted in better saving rate of 3.9% of the ticket price. Although, the saving rate for flight to Hong Kong was improved after 21 days ahead of the departure date, the saving rate remained negative.

Compared to the finding in the study of Domínguez-Menchero et al. (2014), the result in this dissertation was completely different. While the paper of Domínguez-Menchero et al. (2014) indicated that purchasing airline ticket flights 18 days ahead to the departure date was the optimal strategy for most of the experiment routes, the outcome in this paper were different upon each destination. Furthermore, the result that purchasing ticket closer to departure date could provide higher financial benefits for the customers for flight to Hong Kong and Tokyo was unexpected.

TABLE V. EQAUTIONS RESULTED FROM SIMPLIFIED ISOTONIC REGRESSION

Data sets	**Routes**	**Equations**
Data gathered 30 days before departure	Bangkok	$\hat{y} = \begin{cases} -1.99 - 0.28x \text{ if } x \in [-30,-22) \\ 4.23 \qquad \text{if } x \in [-22,-18) \\ 0.93 - 0.18x \text{ if } x\ [-18,-1] \end{cases}$
	Hong Kong	$\hat{y} = \begin{cases} 2.81 + 0.19x \text{ if } x \in [-30,-26) \\ -2.23 \text{ if } x \in [-26,-21) \\ 2.49 + 0.22x \text{ if } x \in [-21,-8) \\ 0.69 \text{ if } x \in [-8,-1] \end{cases}$
	Tokyo	$\hat{y} = \begin{cases} 2.25 + 0.04x\ ifx \in [-30,-22) \\ 1.28\ if\ x \in [-22,-13) \\ 2.99 + 0.13x\ if\ x \in [-13,-10) \\ 1.67\ if\ x \in [-10,-1] \end{cases}$
	Seoul	$\hat{y} = \begin{cases} 3.75 - 0.06x\ if\ x \in [-30,-10) \\ 0.47 - 0.39x\ if\ x \in [-10,-1] \end{cases}$
	Singapore	$\hat{y} = \begin{cases} 12.69 - 0.15x\ if\ x \in [-30,-6) \\ -3.93 - 2.92x\ if\ x \in [-6,-1] \end{cases}$
Data gathered 66 days before departure	Bangkok	$\hat{y} = \begin{cases} 8.33\ if\ x \in [-66,-36) \\ -9.93 - 0.51\mathrm{x}\ if\ x \in [-36,-26) \\ 3.25\ if\ x \in [-26,-2) \\ 0.01 - 1.62\mathrm{x}\ if\ x \in [-2,-1) \end{cases}$
	Hong Kong	$\hat{y} = \begin{cases} -9.68\ if\ x \in [-66,-22) \\ 57.45 + 3.05\mathrm{x}\ if\ x \in [-22,-21) \\ -6.37 + 0.01\mathrm{x}\ if\ x \in [-21,-2) \\ 0.93 + 3.66\mathrm{x}\ if\ x \in [-2,-1] \end{cases}$
	Tokyo	$\hat{y} = \begin{cases} 4.08 + 0.04x\ if\ x \in [-66,-31) \\ 2.70\ if\ x \in [-31,-21) \\ 11.13 + 0.40\mathrm{x}\ if\ x \in [-21,-18) \\ 3.90\ if\ x \in [-18,-1] \end{cases}$
	Seoul	$\hat{y} = \begin{cases} 10.39 \text{ if x } \in [-66,-36) \\ -0.64 - 0.31\mathrm{x} \text{ if x } \in [-36,-1] \end{cases}$
	Singapore	$\hat{y} = \begin{cases} 18.83 \text{ if x } \in [-66,-47) \\ 8.89 - 0.21\mathrm{x} \text{ if x } \in [-47,-9) \\ -0.16 - 1.22\mathrm{x} \text{ if x } \in [-9,-1] \end{cases}$

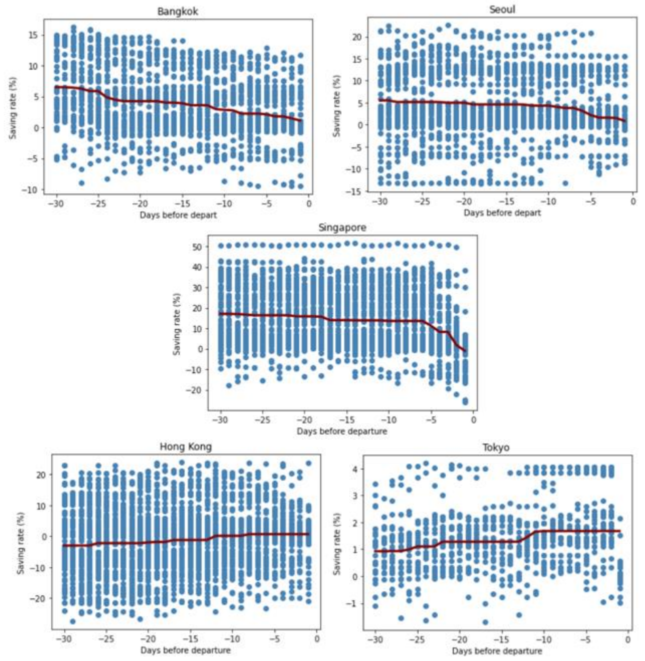

Fig. 5. Isotonic Regression for flight data collected 30 days prior to the departure date.

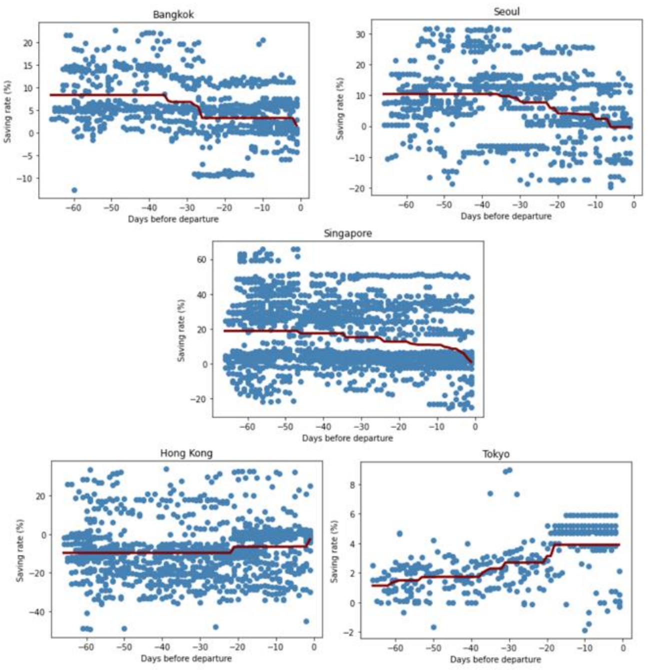

Fig. 6. Isotonic Regression for flight data collected 66 days prior to the departure date.

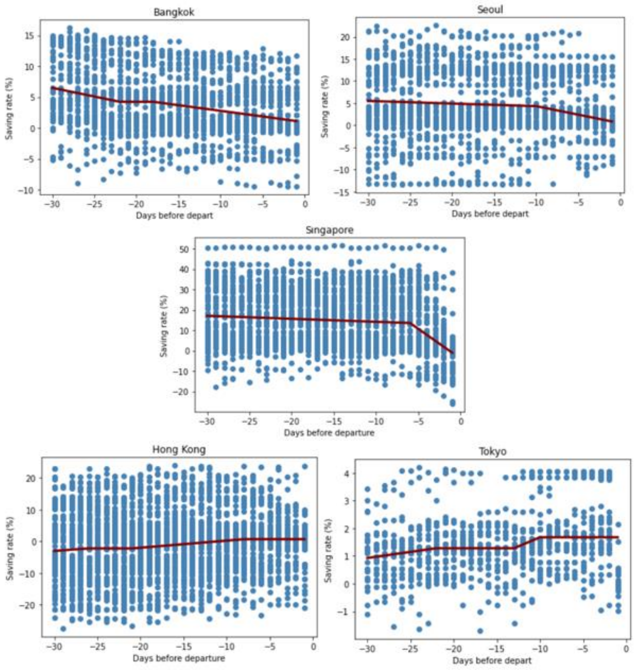

Fig. 7. Simplified Isotonic Regression for flight data gathered 30 days prior to the departure date.

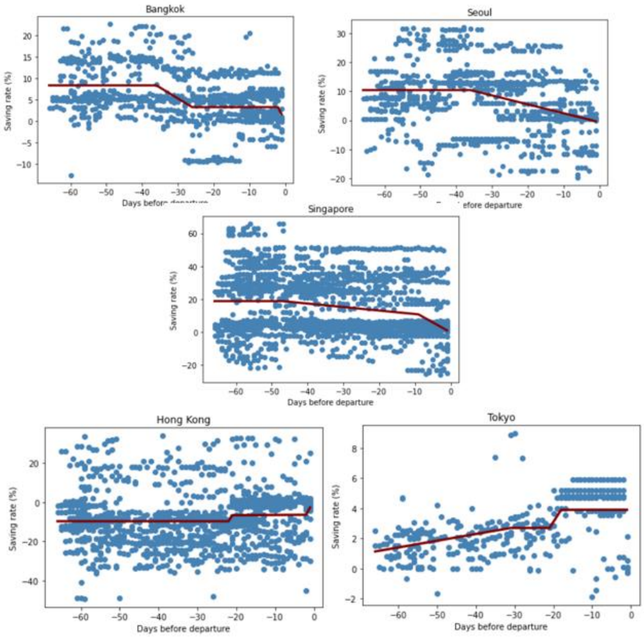

Fig. 8. Simplified Isotonic Regression for flight data gathered 66 days prior to the departure date.

The goodness-of-fit metrics for linear regression and simplified isotonic regression are displayed in Table. VI and Table. VII. Overall, the MSE, MAE and R^2 for linear regression and simplified isotonic regression for each route were not significantly different. The R^2 for all models were close to 0 which indicated the model did not fit well with the data, hence the regression model might not be a good choice to represent this data.

TABLE VI. GOODNESS-OF-FIT RESULTS OF PREDICTED MODEL'S FOR FLIGHT DATA GATHERED 30 DAYS PRIOR TO THE DEPARTURE DATE.

Route	Model	MSE	MAE	R^2
Bangkok	Linear Regression	23.11	3.91	0.09
	Simplified Isotonic Regression	23.1	3.88	0.09
Hong Kong	Linear Regression	93.17	7.54	0.02
	Simplified Isotonic Regression	93.08	7.53	0.02
Tokyo	Linear Regression	1.31	0.87	0.04
	Simplified Isotonic Regression	1.3	0.87	0.05
Seoul	Linear Regression	45.85	5.28	0.03
	Simplified Isotonic Regression	45.45	5.24	0.04
Singapore	Linear Regression	170.56	11.09	0.07
	Simplified Isotonic Regression	163.75	10.84	0.1

TABLE VII. GOODNESS-OF-FIT RESULTS OF PREDICTED MODEL'S FOR FLIGHT DATA GATHERED 66 DAYS PRIOR TO THE DEPARTURE DATE.

Route	Model	MSE	MAE	R^2
Bangkok	Linear Regression	23.11	3.91	0.09
	Simplified Isotonic Regression	23.1	3.88	0.09
Hong Kong	Linear Regression	93.17	7.54	0.02
	Simplified Isotonic Regression	93.08	7.53	0.02
Tokyo	Linear Regression	1.31	0.87	0.04
	Simplified Isotonic Regression	1.3	0.87	0.05
Seoul	Linear Regression	45.85	5.28	0.03
	Simplified Isotonic Regression	45.45	5.24	0.04
Singapore	Linear Regression	170.56	11.09	0.07
	Simplified Isotonic Regression	163.75	10.84	0.1

VI. CONCLUSION

This paper investigated the optimal purchasing time for airline tickets between London to specific Asian capitals by studying the relationship between the number of days before departure and the saving rate if buying a ticket flight on that day. To address this problem, airline ticket flight data were gathered via web scraping algorithm and this data went through preprocessing and regression analysis. Despite the low goodness-of-fit of the model, an isotonic regression could demonstrate the period of days before departure that maximizes the saving rate of airline ticket price in the given routes. For trips to Bangkok, Seoul, and Singapore, early bookings presented the most saving rate, whereas for Hong Kong and Tokyo flights, a few days before departure proved the most advantageous for ticket purchases.

VII. FUTURE WORK

Considerable potential remains for studying airline ticket price behavior and predicting the best time to buy the airline ticket. Adding more factors that impact the airline ticket price such as departure time of the flight, numbers of flight changing, number of airline operate in the specific route (Groves and Gini 2011), or developing more complex prediction model by combining deep learning and natural language processing on social network data to do sentiment analysis for specific event such as concert or conference that might affect the airline ticket price for specific destination (Abdella et al. 2021) could improve the accuracy of the prediction. The concept of saving rate (Domínguez-Menchero et al. 2014) could be applied to finding the best time to purchase similar product to airline tickets such as hotel or tourist packages.

VIII. ACKNOWLEDGMENT

I would like to express my deepest gratitude to my advisor, Dr. Fabrizio Smeraldi, for his guidance and encouragement throughout the process to complete this dissertation. Moreover, I would like to show appreciation to my beloved family and friends for their support.

IX. REFERENCES

Abdella, J.A., Zaki, N.M., Shuaib, K. and Khan, F. (2021). 'Airline ticket price and demand prediction: A survey'. *Journal of King Saud University-Computer and Information Sciences*, 33(4), pp.375-391. Available at: https://www.sciencedirect.com/science/article/pii/S131915781830884X (Accessed: 7 July 2023).

Barlow, R.E., Bartholomew, D.J., Bremner J.M. and Brunk H.D. (1978) *Statistical inference under order restrictions: the theory and application of isotonic regression*. London: John Wiley & Sons. Wiley series in probability and mathematical statistics.

Botchkarev, A. (2019) 'Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology', *Interdisciplinary Journal of Information, Knowledge, and Management*, 14, pp. 45-79. Available at: https://doi.org/10.48550/arXiv.1809.03006

Chicco, D., Warrens M.J., and Jurman G. (2021) 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ Computer Science*, 7. Available at: https://doi.org/10.7717/peerj-cs.623

Domínguez-Menchero, J.S., Rivera, J. and Torres-Manzanera, E. (2014). 'Optimal purchase timing in the airline market'. *Journal of Air Transport Management*, 40, pp.137-143. Available at: https://www.sciencedirect.com/science/article/pii/S0969699714000842?casa_token=bHyJrVHHJ0AAAAAA:HVsd2uqqfSHRok_KJoEENv6nQdSkXhmPi5fu3Uc5VoLiNJ1orajHhYaq-GpIR9xDr8DKJc8WyL8 (Accessed: 7 July 2023).

Etzioni, O., et al. (2003) 'To buy or not to buy: mining airfare data to minimize ticket purchase price', *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 119-128. Available at: https://doi.org/10.1145/956750.956767

Gauthier, T.D. (2001) 'Detecting trends using Spearman's rank correlation coefficient', *Environmental forensics*, 2(4), pp.359-362. Available at: https://www.tandfonline.com/doi/abs/10.1080/713848278 (Accessed: 9 July 2023).

Groves, W. and Gini, M. (2011). *A regression model for predicting optimal purchase timing for airline tickets*. University of Minnesota Digital Conservancy. Available at: https://conservancy.umn.edu/handle/11299/215872 (Accessed: 9 July 2023).

Han, J., Pei, J., and Tong, H. (2023). *Data Mining Concepts and Techniques.* 4th edn. United States: Elsevier.

Khder, M.A. (2021). 'Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application', *International Journal of Advances in Soft Computing & Its Applications*, 13(3), pp.145-168. Available at: https://doi.org/10.15849/IJASCA.211128.11

Kotłowski, W, Koolen, W.M. and Malek, A. (2016) 'Online isotonic regression', *29th Annual Conference on Learning Theory*, United State, 23-26 June. pp.1165-1189. Available at: https://proceedings.mlr.press/v49/kotlowski16.html (Accessed: 9 July 2023).

Kwak, S.K. and Kim, J.H. (2017). 'Statistical data preparation: management of missing values and outliers', *Korean journal of anesthesiology*, 70(4), pp.407-411. Available at: https://doi.org/10.4097/kjae.2017.70.4.407

Maulud, D. and Abdulazeez, A.M. (2020). 'A review on linear regression comprehensive in machine learning', *Journal of Applied Science and Technology Trends*, 1(4), pp.140-147. Available at: https://doi.org/10.38094/jastt1457

NR, R.R. and Vijayalakshmi, M. (2023), 'Web Scrapping Tools and Techniques: A Brief Survey', *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT),* India, 11-12 February. pp.1-4. Available at: https://doi.org/10.1109/ICITIIT57246.2023.10068666

Pedregosa et al. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp.2825-2830 Available at: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (Accessed: 9 July 2023).

Plevris, V., Solorzano, G., Bakas, N.P. and Ben Seghier. (2022) 'Investigation of performance metrics in regression analysis and machine learning-based prediction models', *ECCOMAS Congress 2022: 8th European Congress on Computational Methods in Applied Sciences and Engineering*, Norway, 5-9 June. Available at: https://doi.org/10.23967/eccomas.2022.155

Wen, C.H. and Yeh, Y. (2017). 'Modeling air travelers' choice of flight departure and return dates on long holiday weekends'. *Journal of Air Transport Management*, 65, pp.220-225. Available at: https://www.sciencedirect.com/science/article/pii/S0969699717302831 (Accessed: 6 July 2023).

Wozny, F. (2022). 'The impact of covid-19 on airfares—a machine learning counterfactual analysis'. *Econometrics*, 10(1), p.8. Available at: https://www.mdpi.com/2225-1146/10/1/8 (Accessed: 7 July 2023).

Xu, Y. and Cao, J. (2017). 'OTPS: A decision support service for optimal airfare Ticket Purchase', *2017 IEEE International Conference on Big Data (Big Data)*, United States, 11-14 Dec. pp. 1363-1368. Available at: https://doi.org/10.1109/BigData.2017.8258068