Creating a Flight Prediction Chat-GPT Based Application and Support Vector Machine Algorithm To Better Plan and Price Travel Trips

By Andy Wu

Senior Project

Submitted to the Department of Computer Science & Economics

Yale University

May 2023

Supervised by Alan Weide

DUS: Phillip Strack

Abstract

This paper proposes the development of a full-stack application called Plan.AI for predicting the best time to buy a flight ticket using Support Vector Machines (SVMs). The application will utilize an intelligent recommendation system to analyze historical data and provide predictions for flight prices, allowing users to make informed decisions about when to buy their tickets.

The application will employ SVMs, which are a popular machine learning algorithm used for classification and regression analysis. SVMs have been proven effective in predicting flight prices, making them an ideal choice for this application.

One of the unique features of this application is its ability to analyze historical data to make predictions about future flight prices. By utilizing machine learning algorithms, the application can provide accurate predictions for flight prices, allowing users to make informed decisions about when to buy their tickets.

The application will also provide an interface for users to input their desired travel dates, destination, and other relevant information. Based on this information and historical data, the application will provide a prediction for the best time to buy the ticket to ensure that users get the best possible price.

The high-level overview of the application is that it will act as an intelligent recommendation system for predicting flight prices using SVMs. Users will be able to input their travel information, and the application will provide a prediction for the best time to buy their ticket.

The value proposition of this application is that it will save users money by enabling them to make informed decisions about when to buy their tickets. With accurate predictions for

flight prices, users can avoid overpaying for tickets and ensure that they get the best possible deal.

This paper encourages further research and development of such applications to improve the user experience and provide even more accurate predictions for flight prices.

Introduction

Idea

This paper details the creation of a full stack application for trip planning. The application will analyze inputs from users and then recommend choices for the user to easily make decisions for hotels, planes, restaurants, and travel. For example as you're typing in the location, you will be able to see recommendations of popular destinations for vacation in that area.

This application will be able to analyze keywords from what you're typing. When you're typing a word like hotel, you'll see recommended hotels in that area that you will have the option to book or to not book. Once you've made a decision, the links to these websites will be automatically hyperlinked in the data. At the bottom of the application, you'll have an approximate cost of the trip, which aggregates costs from each aspect of the trip. You will also be able to add users to this trip planning document, allowing you to easily split costs.

Purpose

Airline ticket prices are notorious for being unpredictable and constantly fluctuating. This can make it difficult for travelers to plan and budget their trips accordingly. To address this issue, various online travel booking platforms such as Google Flights and Kayak have implemented flight prediction algorithms to help users make more informed decisions about when to book their flights.

However, these platforms often use black-box algorithms, leaving users uncertain about how the predictions are generated. This lack of transparency can limit users' trust in the prediction results and prevent them from fully utilizing the predictive power of the algorithms.

In this paper, we propose a new approach to flight price prediction that utilizes support vector machines (SVMs) to model the complex relationships between ticket prices and various factors such as departure date, flight duration, and route popularity. Unlike the proprietary algorithms used by major travel booking platforms, the SVM-based model provides users with a clear and interpretable way to understand how flight price predictions are generated.
By implementing the flight prediction model, users will be able to make more informed decisions about when to book their flights, ultimately saving time and money. Furthermore, the approach highlights the importance of incorporating mathematical models and transparent algorithms into the travel industry, offering a new perspective on how machine learning can benefit both consumers and service providers in this domain.

Breakdown of Paper

The methodology section of the paper provides a detailed description of the training data, including the data collection process and data cleaning procedures. The paper then outlines the feature selection process and the machine learning algorithms used to develop the predictive model. The selected algorithms are evaluated using cross-validation techniques and hyperparameter tuning to optimize performance.

The results reveal the data and output from the model and application. The discussion section of the paper provides an interpretation of the results and discusses the implications of the research. The paper concludes by highlighting the significance of the research and its potential applications in the travel industry.

High Level Overview of the Application

This application will act as an intuitive trip planning document interface that will centralize decision making and trip planning information in one place.

Value Proposition

Users will be able to save hours and days of their time. Currently, people spend a lot of time planning for trips because they have to navigate through dozens of different trip planning sites. This application allows you to quickly make decisions as you are planning the trip.

Features

- Trip Planning Document
  - Input Boxes
  - Recommendation boxes
- Travel Algorithm
  - Pulls in existing API data and parses them to provide optimized travel pricing/choices
- Input Analysis
  - Analyze keywords from people’s inputs → give recommendations

Background

Literature Review

The airline industry is a highly competitive market where pricing strategies can greatly impact an airline's profitability. Predicting airline ticket prices and demand has been an ongoing challenge for airlines, and recent advances in machine learning have opened new opportunities to improve pricing strategies. Various studies have identified key features that are utilized in predicting airline ticket prices and demand, such as the length of travel, source, destination, past prices, brand loyalty, date of travel, and planning ahead (Abdella et al., 2019).

However, airlines often do not provide publicly available data, and alternative data sources such as sentiment analysis, web scraping, and social media must be utilized to obtain the necessary data (Biswas et al., 2022). Biswas et al. (2018) conducted a survey on the use of machine learning in predicting airline ticket prices and demand, and found that several machine learning algorithms were used for prediction, including support vector regression, neural networks, and decision trees.

One popular machine learning technique used for predicting airline ticket prices and demand is XGBoost, which is a gradient boosting algorithm that uses decision trees as base learners (Sarao 2022). XGBoost has been used in various studies to predict airline ticket prices and demand, such as in the case study by Sarao (2022). The study utilized XGBoost along with other machine learning algorithms, such as K-nearest neighbor and linear regression, to predict flight prices. The results showed that XGBoost was the most accurate algorithm, achieving an accuracy of 86.63% compared to other algorithms, which had accuracies ranging from 75.55% to 80.76%.

In addition to predicting airline ticket prices, machine learning has also been applied to the hotel industry to optimize dynamic pricing. Rust and Cho (2018) utilized machine learning techniques such as decision trees, random forests, bootstrapping, bagging, and XGBoost to predict hotel prices dynamically. The results showed that XGBoost outperformed other algorithms in terms of accuracy and speed.

Despite the promising results of utilizing machine learning to predict airline ticket prices and demand, there are still limitations and challenges to consider. One major limitation is the lack of publicly available data from airlines, which requires the use of alternative data sources that may not always be accurate or comprehensive. Another challenge is the need for frequent updates and retraining of machine learning models to ensure that they remain accurate and relevant over time.

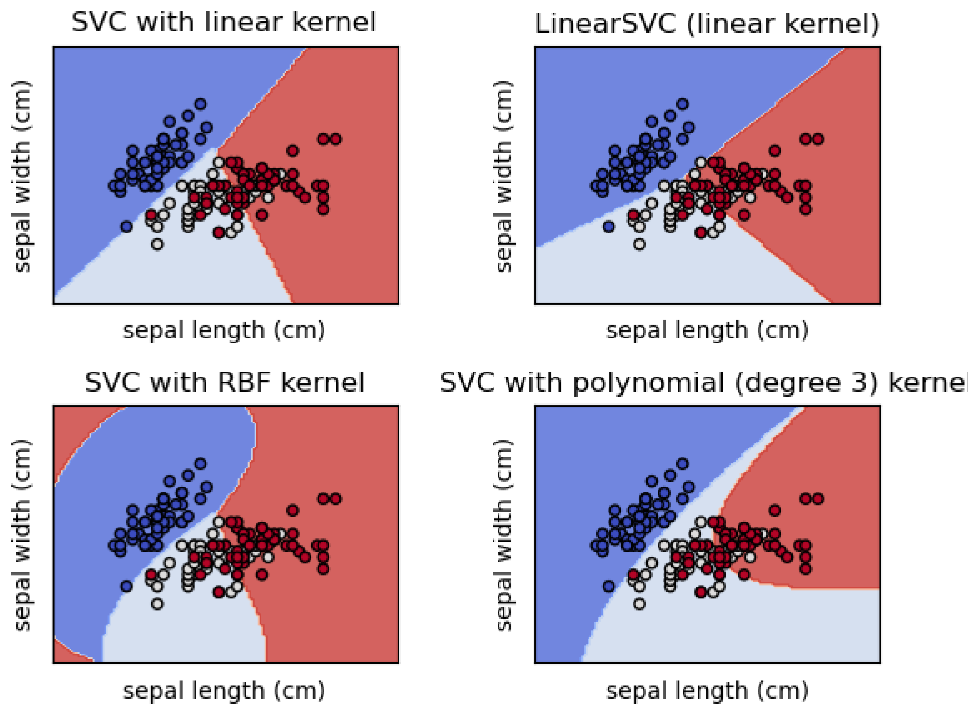
Overall, machine learning has proven to be a powerful tool in predicting airline ticket and hotel prices, with various algorithms utilized to achieve the highest accuracy. Future research may consider additional features and alternative data sources to further improve the accuracy of these predictions.

SVM Training

Support Vector Machines (SVM) are a type of machine learning algorithm used for classification and regression analysis. SVMs are particularly useful when dealing with complex data that has multiple variables, as they can help identify the underlying patterns in the data and use them to make accurate predictions.

SVMs work by finding the best possible line or hyperplane that separates data points of different classes in a high-dimensional space. The hyperplane is chosen so that it maximizes the margin between the two classes, i.e., the distance between the hyperplane and the closest data points of each class. SVMs can also use different kernel functions to transform data into a higher-dimensional space where the data is more easily separable.

SVMs have a wide range of applications, including image and speech recognition, natural language processing, and bioinformatics. They are also commonly used in finance for predicting stock prices and in marketing for customer segmentation. SVMs have proven to be very effective in many real-world scenarios, especially when dealing with large datasets and complex data structures. Specifically for the application, it’s useful for best determining and segmenting the flight data.

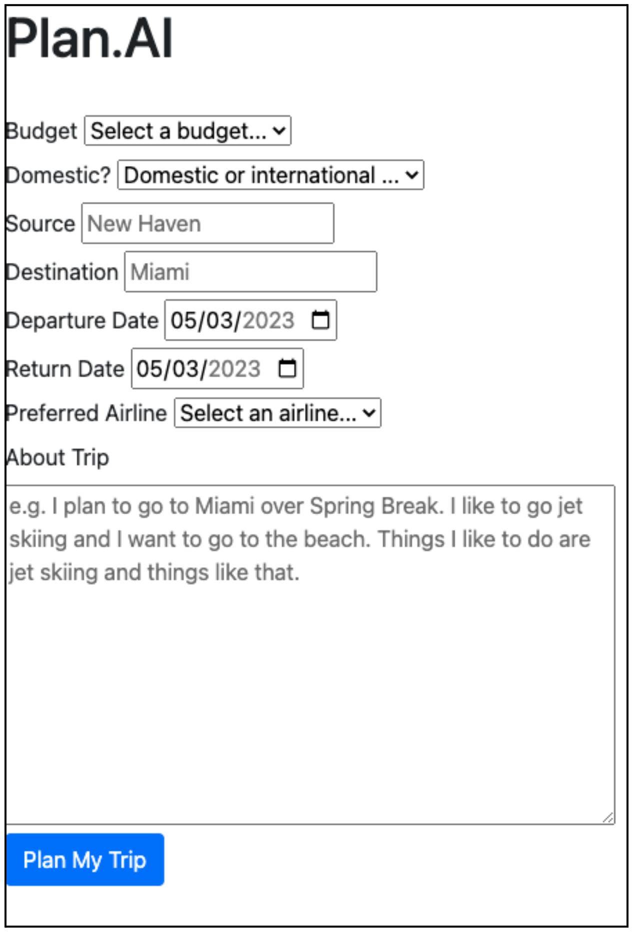


Example of How Support Vector Machines Work

(https://scikit-learn.org/stable/modules/svm.html)

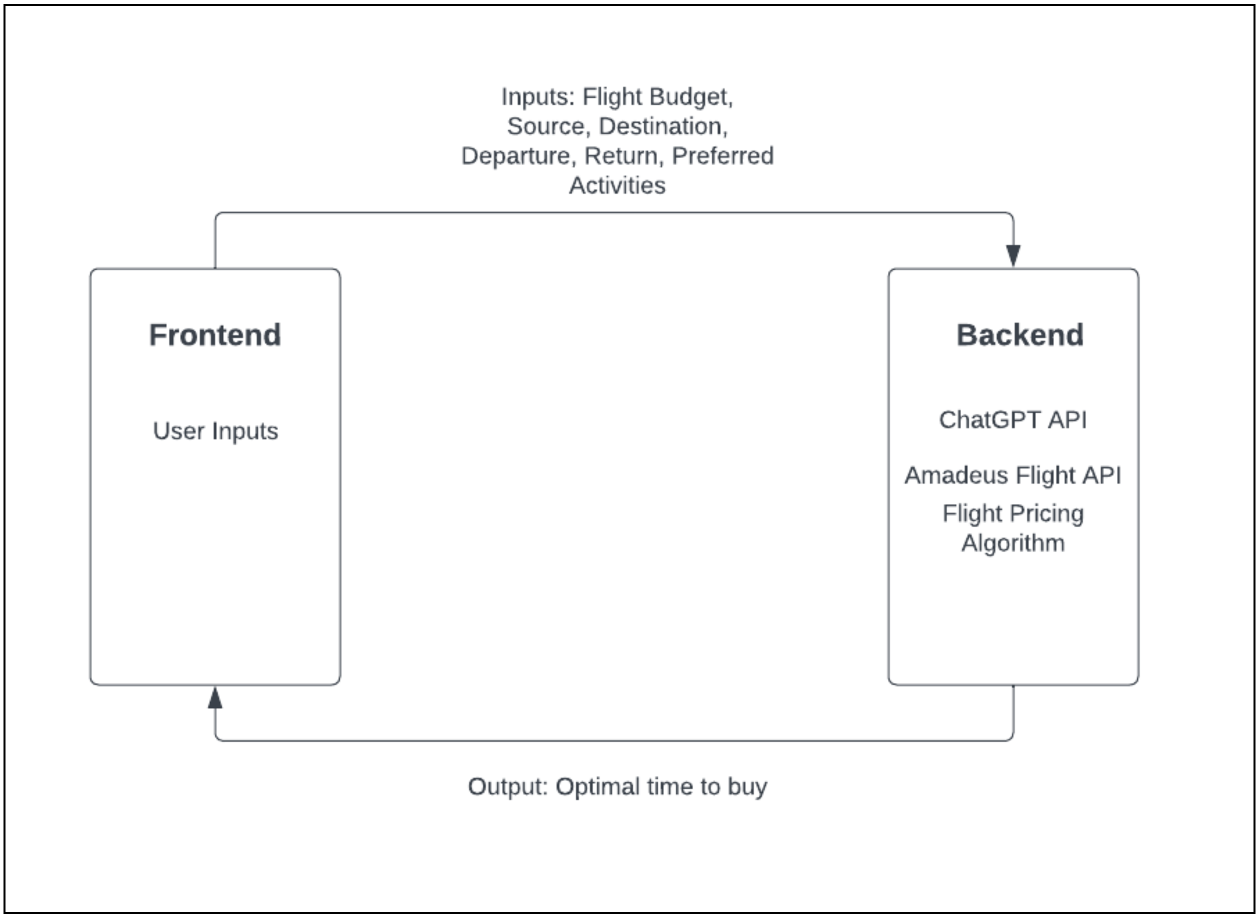
Methodology

Introduction



Screenshot of the Application

The application was built using a full stack framework. The frontend is responsible for collecting user inputs and displaying the results returned by the backend. The backend of this application is responsible for processing the user inputs and then returning the results to the frontend. The frontend and backend work together to provide a seamless user experience to better plan their travel.



Application Flowchart

Frontend

The application is built using the Flask framework. The frontend was designed in HTML and CSS with Python for the backend. It is designed to take in data from the user about their dates of trip, preferred flights, budgets, and what they like to do when they are on trips to then better plan their trip while on that trip. The application also uses the ChatGPT API in the backend to help users plan activities while on their trip and then recommends things to do in the place they're going to.

Backend

One of the key features of the application is the pricing algorithm in the backend which takes people's inputs. Using a Python Jupyter notebook, the application trains a support vector machine model on people's data to predict a price based on the days before the flight, duration of the flight, and the place of the flight. The model is trained using a flight dataset from Kaggle (linked in sources). The model's training parameters include flight time, destination, and days_left, with the hypothesis that flight time will result in more gas usage, destination will determine prices because more popular places will have higher prices, and flight duration will also impact prices. The pre-trained model is saved using Pickle and then loaded when the user gives inputs.

Our backend uses Application Programming Interfaces (APIs) provided by Amadeus and ChatGPT. Amadeus provides us with the capability to return flight offerings with the given parameters from the user, while ChatGPT is used to generate possible activities for users based on their preferences.

In addition, we have developed several key programs to enhance the functionality of the backend. For example, we have implemented a distance finder between cities to enable the application to return more accurate results based on the user's location. We have also developed a ChatGPT Prompt Generator to generate prompts for ChatGPT based on the user's inputs, which enables us to provide more personalized recommendations.

Overall, the methodology involves a combination of front-end and back-end development using established frameworks and APIs, as well as custom-developed programs to enhance the functionality and personalization of the application.

SVM Flight Pricing Algorithm

Now, let's see how we can train an SVM model on 4 parameters, namely Source City, Destination City, Duration, and Days Till Flight. In this scenario, we will be using one hot encoding for the categorical variables, i.e., Source City and Destination City. We will also be using the geopy library to calculate the distance and duration between the two cities.

Here are the steps followed to train the SVM model:

1. Preprocessing the data:

Start by preprocessing the data. Then convert the categorical variables, i.e., Source City and Destination City, into binary variables using one hot encoding. Calculate the distance and duration between the two cities using the geopy library. Finally, split the data into training and testing sets.

2. Training the SVM model:

Once the data is preprocessed, train the SVM model on the training data. Use the SVM implementation provided by the scikit-learn library. Choose the appropriate kernel function for our model. The most common kernel functions used are linear, polynomial, and radial basis function (RBF). Experiment with different kernel functions and choose the one that performs the best on our dataset.

3. Evaluating the SVM model:

After training the model, evaluate its performance on the testing data. Use metrics such as accuracy, precision, recall, and F1-score to evaluate the performance of our model.

4. Visualizing the SVM model:

Finally, visualize the SVM model using matplotlib. Plot the support vectors, the decision boundary, and the margins between the two classes.

In summary, the SVM model was trained on 4 parameters, namely Source City, Destination City, Duration, and Days Till Flight, using one hot encoding for the categorical variables and the geopy library for distance and duration calculations. The scikit-learn library provided an implementation of SVM used for the model. The model is saved using pickle and then loaded in the application to evaluate on user data.

Results

Plan My Trip

Here is the Trip Planning Data: Here are possible options of things you can do: 1. Take a jog along the Miami Beach Boardwalk. The scenic route spans 4 miles and is perfect for a morning or evening run. 2. Join a local running club. Miami offers various running groups, such as the Miami Runners Club and the Brickell Run Club, which organize weekly runs and events. 3. Explore the city's parks and trails. Miami has several parks and nature reserves, includin the Oleta River State Park and the Amelia Earhart Park, that offer scenic trails for runners. 4. Attend the Miami Book Fair. The annual event takes place in November and features author readings, book signings, and literary discussions. 5. Visit local bookstores. Miami has several independent bookstores, such as Books & Books and the Miami Beach Books & Books, that offer a vast selection of books and host author events. 6. Take a guided literary tour. The Miami Literary Tour takes visitors on a journey through the city's literary history, showcasing famous authors and landmarks. 7. Attend book clubs an literary events. Many local libraries and bookstores host book clubs and literary events, providing opportunities to connect with other book lovers and discuss literature. Here are the prices for these options and you can book nowHere is what the expected prices look like in the next week

ChatGPT Recommendations in response to “I like to run and read”

Data

	Unnamed: 0	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	0	4	1408	2	2	2	5	5	1	2.17	1	5953
1	1	4	1387	2	1	2	4	5	1	2.33	1	5953
2	2	0	1213	2	1	2	1	5	1	2.17	1	5956
3	3	5	1559	2	4	2	0	5	1	2.25	1	5955
4	4	5	1549	2	4	2	4	5	1	2.33	1	5955
...	...	...	...	...	...	...	...	...	...	...	...	...
300148	300148	5	1477	1	4	0	2	3	0	10.08	49	69265
300149	300149	5	1481	1	0	0	5	3	0	10.42	49	77105
300150	300150	5	1486	1	1	0	5	3	0	13.83	49	79099
300151	300151	5	1483	1	1	0	2	3	0	10.00	49	81585
300152	300152	5	1477	1	4	0	2	3	0	10.08	49	81585

Training Dataset Used (Kaggle)

	source_city	destination_city	duration	days_left
27131	2	4	19.75	40
266857	4	5	9.83	42
141228	4	0	10.50	41
288329	1	2	14.50	14
97334	0	5	8.25	20
...	...	...	...	...
5234	2	5	12.58	27
5591	2	5	2.17	29
168314	3	0	13.58	28
175191	3	4	2.00	40
287693	1	2	25.42	3

Training Dataset After Sampling, Cleaning, and One-hot Encoding (Kaggle)

H23M', 'id': '35', 'numberOfStops': 0, 'blacklistedInEU': False}]}], 'price': {'currency': 'EUR', 'total': '399.09', 'b
ase': '346.00', 'fees': [{'amount': '0.00', 'type': 'SUPPLIER'}, {'amount': '0.00', 'type': 'TICKETING'}], 'grandTotal'
: '399.09'}, 'pricingOptions': {'fareType': ['PUBLISHED'], 'includedCheckedBagsOnly': False}, 'validatingAirlineCodes':
 ['AS'], 'travelerPricings': [{'travelerId': '1', 'fareOption': 'STANDARD', 'travelerType': 'ADULT', 'price': {'currenc
y': 'EUR', 'total': '399.09', 'base': '346.00'}, 'fareDetailsBySegment': [{'segmentId': '23', 'cabin': 'ECONOMY', 'fare
Basis': 'NH3OAVMN', 'brandedFare': 'MA', 'class': 'N', 'includedCheckedBags': {'quantity': 0}}, {'segmentId': '35', 'ca
bin': 'ECONOMY', 'fareBasis': 'QAA0JKEN', 'brandedFare': 'ECONOMY', 'class': 'Q', 'includedCheckedBags': {'quantity': 0
}}]}]}, {'type': 'flight-offer', 'id': '249', 'source': 'GDS', 'instantTicketingRequired': False, 'nonHomogeneous': Fal
se, 'oneWay': False, 'lastTicketingDate': '2023-05-05', 'lastTicketingDateTime': '2023-05-05', 'numberOfBookableSeats':
 5, 'itineraries': [{'duration': 'PT1H29M', 'segments': [{'departure': {'iataCode': 'LAX', 'terminal': '6', 'at': '2023
-05-09T07:00:00'}, 'arrival': {'iataCode': 'SFO', 'terminal': '2', 'at': '2023-05-09T08:29:00'}, 'carrierCode': 'AS', '
number': '3471', 'aircraft': {'code': 'E75'}, 'duration': 'PT1H29M', 'id': '23', 'numberOfStops': 0, 'blacklistedInEU':
 False}]}, {'duration': 'PT1H24M', 'segments': [{'departure': {'iataCode': 'SFO', 'terminal': '3', 'at': '2023-05-12T08
:38:00'}, 'arrival': {'iataCode': 'ONT', 'terminal': '2', 'at': '2023-05-12T10:02:00'}, 'carrierCode': 'UA', 'number':
'5740', 'aircraft': {'code': 'CR7'}, 'duration': 'PT1H24M', 'id': '36', 'numberOfStops': 0, 'blacklistedInEU': False}]}
], 'price': {'currency': 'EUR', 'total': '399.09', 'base': '346.00', 'fees': [{'amount': '0.00', 'type': 'SUPPLIER'}, {
'amount': '0.00', 'type': 'TICKETING'}], 'grandTotal': '399.09'}, 'pricingOptions': {'fareType': ['PUBLISHED'], 'includ
edCheckedBagsOnly': False}, 'validatingAirlineCodes': ['AS'], 'travelerPricings': [{'travelerId': '1', 'fareOption': 'S
TANDARD', 'travelerType': 'ADULT', 'price': {'currency': 'EUR', 'total': '399.09', 'base': '346.00'}, 'fareDetailsBySeg
ment': [{'segmentId': '23', 'cabin': 'ECONOMY', 'fareBasis': 'NH3OAVMN', 'brandedFare': 'MA', 'class': 'N', 'includedCh
eckedBags': {'quantity': 0}}, {'segmentId': '36', 'cabin': 'ECONOMY', 'fareBasis': 'QAA0JKEN', 'brandedFare': 'ECONOMY'
, 'class': 'Q', 'includedCheckedBags': {'quantity': 0}}]}]}]

Example Output Returned by Amadeus API for LAX to SFO. Includes, prices, checked bags status, and currency type

```
In [31]: y_pred

Out[31]: array([ 6568,   6568,  55983,   6568,   6568,   6568,   5955,  54608,
               55983,   6568,  51595,   6568,  44280,   6450,   4499,   6568,
                6568,   6568,   2339,   6568,   6568,  49613,   6568,   6568,
                5954,   6568,   6568,   6568,   6568,  55377,   6568,   6568,
                6568,   6568,   6568,   6568,   6568,   6568,   6568,   6568,
               51595,  60260,   6568,   6568,   6568,   6568,   6568,   6568,
                6568,  54608,   6568,   6568,   5954,   6568,   6568,   6568,
                7424,  60260,   6568,   4416,   6568,  57920,   6568,  60260,
               13830,  54608,   6568,   6568,   6568,   6568,  54608,  54608,
                6568,   6568,   6568,   6568,   6568,   6568,   6568,   5954,
                6568,   6568,   6568,   6568,  54608,  51595,   6568,   6568,
                2050,  23838,   4755,   5177,   6568,  13830,   6568,   6568,
               60260,   6568,  54608,  54608,   6568,   6568,  49613,   6568,
                2050,   6568,  28510,   6568,   6568,   6568,   6568,   6568,
                6568,   6568,   6568,   6568,   6568,   6568,  56588,   6568,
               54684,  60260,   6568,  56588,   6568,   6568,  42521,   6568,
                6568,   6568,   6568,  54608,   6568,   6568,   6568,   6568,
               49725,   6568,  60260,   6568,   6568,   6568,  11205,   6568,
                5960,   6568,   2050,  52063,   6568,   6568,   6568,   6568,
```

Output returned during price classification period for eval dataset

Discussion

The application discussed in this research paper is a promising example of how machine learning and APIs can be used to improve travel planning. ChatGPT is able to provide helpful recommendations to the user to plan the trip around their interests during their stay.

By leveraging a Kaggle dataset and an SVM model, the application is able to help better navigate flight prices. The SVM model is trained using one-hot encoding, allowing it to predict flight prices based on several features such as the user's source, destination, duration, and day until the flight. Once the user inputs their travel details, the model predicts the expected price for the trip, which the user can then compare to the available options retrieved through the Amadeus API. This enables the user to make an informed decision about whether to book now or wait for prices to potentially decrease.

The results of this application are promising for travel planning. By providing data-driven recommendations, users can make more confident decisions about their travel plans, particularly those who are on a tight budget. They can plan their trips around optimal times to save money on flights and other travel-related expenses. Frequent travelers can also benefit from this application by waiting for the best deals.

The significance of this research lies in its potential to improve the travel industry. The application serves as a proof-of-concept for the integration of machine learning and APIs into travel planning. As machine learning algorithms continue to improve, we can expect to see more innovative applications like this in the travel industry. This research provides a foundation for future research on how machine learning can be leveraged to improve the travel experience for individuals and families.

Conclusion

This research highlights the potential of machine learning and APIs to improve travel planning. The application discussed demonstrates the value of data-driven recommendations and the potential to help users save money and time on travel-related expenses. As the travel industry continues to evolve, the integration of technology into travel planning will play an increasingly important role.

The findings of this research show that the developed application can provide personalized recommendations to users based on their preferences and budget. Additionally, the SVM-based pricing algorithm can generate accurate predictions of flight prices and recommend the best time for users to buy tickets. This research is significant because it provides insight into the development of personalized travel planning applications and the use of machine learning algorithms for pricing predictions.

This senior thesis paper presents the design and development process of a front-end application for personalized travel planning. The use of Flask, HTML, CSS, and Javascript, along with ChatGPT API and SVM-based pricing algorithm, demonstrates the effectiveness of integrating various technologies for developing applications that cater to user preferences. This research provides insights for further development of personalized travel planning applications and the use of machine learning algorithms in the travel industry.

Future Directions

One potential area of future development is exploring the use of neural networks in the travel planning app. Neural networks are a type of artificial intelligence that can be used to recognize patterns in data and make predictions based on those patterns. In the context of travel planning, neural networks could be used to further personalize recommendations for users based on their travel history and preferences.

Another potential area of development is conducting more user testing to determine the practicality and usability of the travel planning app. While the app may save users time, it's important to determine whether or not users find the app useful and whether they would be willing to use it in practice. User testing can also help identify any areas where the app could be improved or streamlined to make it more user-friendly.

Additionally, the app could be expanded to include more features and functionality beyond just flight recommendations and travel activities. For example, the app could integrate with hotel booking systems or provide information on local transportation options. By providing a more comprehensive suite of travel planning tools, the app could become a one-stop-shop for all travel planning needs.

Overall, the future directions of the travel planning app and support vector machines include exploring the possibility of neural networks, conducting more user testing to determine the practicality of the app, and expanding the app's functionality beyond just flights and travel activities. By continuing to develop and refine the app, it has the potential to save users time and provide a more personalized and streamlined travel planning experience.

References

1. Abdella, Juhar Ahmed, et al. “Airline Ticket Price and Demand Prediction: A Survey.” Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 4, 2021, pp. 375–391, https://doi.org/10.1016/j.jksuci.2019.02.001.
2. John Rust & Sungjin Cho, 2018. "Optimal Dynamic Hotel Pricing," 2018 Meeting Papers 179, Society for Economic Dynamics. <https://ideas.repec.org/p/red/sed018/179.html>
3. Abdella, Juhar Ahmed, et al. “Airline Ticket Price and Demand Prediction: A Survey.” Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 4, 2021, pp. 375–391, https://doi.org/10.1016/j.jksuci.2019.02.001.
4. Ijraset. “IJRASET Journal for Research in Applied Science and Engineering Technology.” Flight Price Prediction: A Case Study, www.ijraset.com/research-paper/flight-price-prediction. Accessed 3 May 2023.
5. Sarao, Parwaz, and Pushpendu Samanta. “Flight Fare Prediction Using Machine Learning.” SSRN, 7 Dec. 2022, papers.ssrn.com/sol3/papers.cfm?abstract_id=4269263.
6. Engel, Andrew. “Categorical Variables for Machine Learning Algorithms.” Medium, 16 Mar. 2022, towardsdatascience.com/categorical-variables-for-machine-learning-algorithms-d2768d587ab6.
7. Bathwal, Shubham. “Flight Price Prediction.” Kaggle, 25 Feb. 2022, www.kaggle.com/datasets/shubhambathwal/flight-price-prediction.
8. Gandhi, Rohith. “Support Vector Machine - Introduction to Machine Learning Algorithms.” Medium, 5 July 2018,

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

9. “Support Vector Machines”. https://scikit-learn.org/stable/modules/svm.html
10. “CSEC Senior Project.” GitHub, https://github.com/andywu8/submission-csec-finalproject. Accessed 3 May 2023.