DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

# FLIGHT TICKET PURCHASE TIME PREDICTION

PROJECT REPORT
CSCI – 6505 MACHINE LEARNING

GROUP NAME: TEAM 15


| | |
|---|---|
| DHRUV PINTUBHAI DOSHI | B00883311 |
| VISHAL RAKESH JAISWAL | B00867181 |
| KISHAN RAKESHBHAI PATEL | B00882970 |
| PATHIK KUMAR PATEL | B00869765 |

# TABLE OF CONTENTS

## ABSTRACT

The growing demand in the aviation industry is pumping flight prices up. For an end-user, it is difficult to determine whether they should book the ticket right away or wait for a better price to come for the ticket. Often, someone waits for the better price, and the ticket price gets more and more increased day by day. Based on the historical data, several models could predict the flight's future price. They do not entirely answer the question; hence, this project targets that specific issue. There are three proposed models based on machine learning to better answer whether the user should buy the ticket right away or wait for the prices to fall. CNN, SVM, and Q Learning are used on top of the Baseline model of the Dummy classifier. The research suggests that the SVM implementation is better than the CNN one, and Q-Learning could be the potential breakthrough clubbed with DQN and DDQN algorithms.

**Keywords**: Machine Learning; Algorithms; SVM; CNN; Q-Learning

## INTRODUCTION

Almost all airlines use a complex algorithm to adjust the flight rates majorly according to the time of the year. It also includes seat availability, oil prices, airline capacity, and current rules and regulations. Many attributes contribute to deriving the price of the flight tickets; hence, the customer willing to purchase a ticket had either one option of choosing the costly ticket or waiting for the reduction of price in the future, which is not guaranteed. Customers who are not familiar with these algorithms usually pay more than the actual cost for purchasing the ticket as soon as possible.

However, there is an enormous amount of data available in the context of flight ticket prices, but in the project, we have used the data to help the customers make their decisions more swiftly. We are creating an online ticket buying agent who would help customers buy the tickets at optimal prices and offer the best scenario in the context of buying tickets.

The input for our algorithm would question whether the customer should buy the ticket at a given price for some specific date, and the model would return the answer in terms of either going ahead with the purchase, or they can wait for another opportunity or re-evaluate whether to make a purchase or not.

## LITERATURE SURVEY

Going in this domain, there had not been a massive amount of previous work done in this direction; most of the work done with machine learning in this segment was regarding the price prediction of the future dates with the historical data. This project aimed to determine whether the user should move ahead with the ticket or wait to get better pricing.

To determine the quarterly average airline price fare, [1] suggested that based on public datasets (DB1B and T-100) using a machine learning framework, they neglected the several features of the datasets and filtered out the required parts accordingly. Using this model, they could attain an adjusted R squared score of 0.896. This paper inherits the outcome of other papers which dealt with regression stacking [2] and linear units to improve the restricted Boltzmann machine [3]. Moving ahead, they found that the support vector machine and random forest were most frequently used in predicting the airfare price [4][5][6]. Another aspect critical for the working of the model

was demand and supply equilibrium for the flight tickets as if there is a surge in demand the supply is stagnant, then this results in a higher amount of price, and this works vice versa also [7]. Dealing with this was [8] handled with data-driven modeling and data-driven frequency-based profit maximization [9].

It is hard for the client to buy an airline ticket at the most reduced cost. To determine the ideal purchase time for flight tickets, Gini and Groves [10] exploited Partial Least Square Regression (PLSR) for building up a model. Continuing that, an investigation from Dominguez-Menchero suggests that [11] the perfect time for buying the airline ticket depends on the nonparametric isotonic backslide technique for the specific ticket. The model works around the adequate number of days before buying the ticket [12].

Now, how machine learning can be applied to the time series problem could be further determined by implementing the flight price prediction as to the base problem and making a model that suggests to the user whether they should move ahead with the purchase or not [13]. Further on, using the Unsupervised Machine learning algorithms could help advance in this direction [14].

The basic idea of the prediction is to develop the model which could work on the face of Lazy prediction with the Regression Analysis, [15] alongside that a regression-based schema could be used and applied to the data to predict the prices, [16] this schema allows us to have a network which could suggest the alternative paths if required.

## PROBLEM STATEMENT

After looking through the literature survey, it is evident that there are multiple models and projects which could predict the price of the flight ticket in the upcoming days, but there was not a single model which answers the user whether they should move ahead with the purchase, or they should wait for the better price in the future. Hence this is the problem statement which we are targeting.

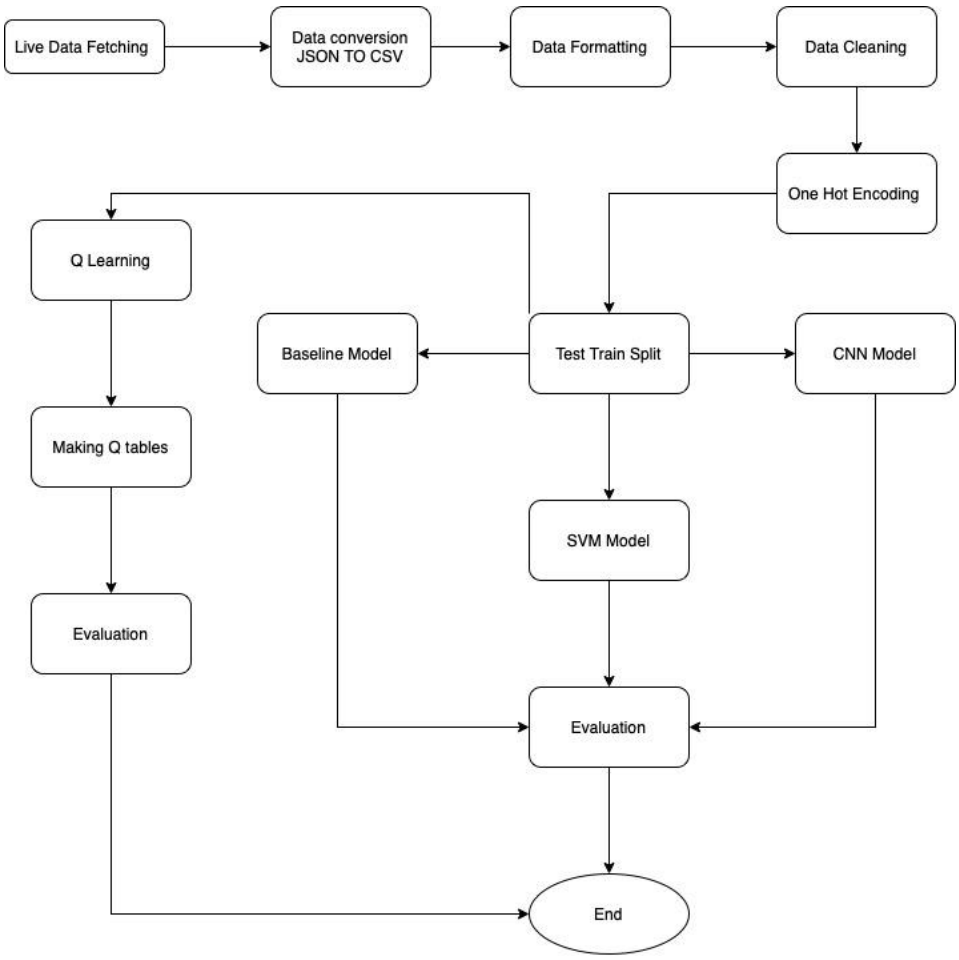
To answer that, whether it is a wise decision to buy the ticket right now at the given price or the user should wait for the price to fall soon and then get the same ticket at the lower cost.

Current algorithms and the approach for this scenario do not solve the use case for giving clarity to the user regarding whether they should buy the ticket or not; the current models suggest the price in the future concerning the previous trends and demand. Now, this does not give the user the confidence whether they would be able to secure the ticket in the future with the price which the model is portraying.

The previous research uses the dataset fetched from the previous flights, and they somehow lack relevance with the current situation, whereas, for this specific issue, the live data must be fetched.

## PROPOSED TECHNIQUE

For the development of multiple models, a flowchart is given below, which gives an overview of the flow of the project. Baseline, SVM, and CNN models are supervised models. Hence, they are compared with the same evaluation matrix, and Q-Learning is based on a reinforcement learning algorithm hence the evaluation matrix for that model is different compared to other models.




### DATA COLLECTION

As mentioned before, there was a need for the live data to be fetched. Several APIs provided solutions for direct data collection like Google QPX, Skyscanner, Aero Databox, and Travel Payouts. However, most of them were either down or restricted by the admin; hence we settled for Travel payouts API for data collection.

This API is used from RapidAPI, and the only back draw of this API was that it gets updated every 48 hours. We developed a script that takes the data from the API and saves it in JSON format to resolve this issue. This script also makes sure to be relevant with the latency of the API to remove the null entries where the second call is triggered before the acceptance of the API.

The collected data was in JSON format and converting that directly to CSV was not possible; hence one more script was written in context to convert the data from JSON to CSV, making it more acceptable with Pandas, a library of Python.

**DATA VISUALIZATION**

For data visualization, the seaborn python library is used with pair plot and heatmap graphics, which helps know data in-depth and factor the important columns and data relevant to the model training. Furthermore, dimension reduction is also based on these plots and maps for efficient training and a better model.

**DATA PRE-PROCESSING**

The data fetched and saved in CSV is pretty much cleaned as the data was fetched through the API and having that latency adjustment in the scripts removed all the instances where the null values could occur.

**DATA CLEANING**

The API fetched the data live from the different agents; hence, there were many duplicate data. Along with that, from the visualization, multiple columns were not relevant to the use case.
Construction of new columns required for training of data alongside that construction labels like Buy and Wait.

1. DATA FORMATING -DATE

   The data format is pandas' object; hence it is converted to pandas date-time. Added one more column named days to departure from the data, which was done after transforming pandas object to panda's date-time.

2. ONE HOT ENCODING

   The data had information that was difficult to feed in the model; hence, one hot encoding was done on all the data converting most of them to either 0 or 1.

**TRAIN TEST SPLIT**

For the development of this project, we have taken the ratio of 4:1 for training and testing, aggregating 80% for training and 20% for testing the model.

**MODELS**

1. BASELINE MODEL

A dummy classifier creates a baseline model. This simple model would ensure that the customer always gets the ticket, and the tickets will not get sold out.

Dummy classifier will not take trends and identifiers are not considered, and it only uses the simple rules to predict the class labels. Thus, it is the simple baseline for other classification algorithms. Thus, another machine learning model is expected to perform better than the baseline model.

2. CNN

Convolution Neural networks are used to develop the supervised model, and the reason for using it as the first choice is because of parameter sharing and dimensionality reduction embedded in the concept.

Keras Python library is being used to develop multiple embedded layers in CNN. We have converted one-dimensional data according to the input matrix format suitable for CNN.

3. SVM

Support Vector Machines are used as they are supervised models, and the reason for using them over others is better effectiveness for higher dimensional spaces and the memory efficiency because of using a subset of training points in decision function.

SVC method of SVM is used, and for the Kernel, RBF kernel is used.

4. Q-LEARNING

Q learning is reinforming a learning algorithm that involves the construction of Q tables based on the reward function. In this case, the reason for using it is that the agent has a choice of actions to make between buying and waiting; after making an action, it calculates the reward and moves ahead to the next stage.

The q values are calculated until a terminal state is reached and the agent can no longer find the new Q values. Here the goal is to maximize the rewards.

The Q table has an array of actions specifying the Q values.

The equation for evaluating Q values,

$$Q(a, s) = R(s, a) + \gamma max_{a'}(Q(a', s'))$$

Here, s denotes state, action, and $\gamma$ denotes a discount factor.

The equation for reward function,

$R(s, buy) = -price(s)$
$R(s, wait) = -Q;\ Penalty\ if\ flights\ sell\ out$
$Else\ R(s, wait) = 0$

## EVALUATION METRICS

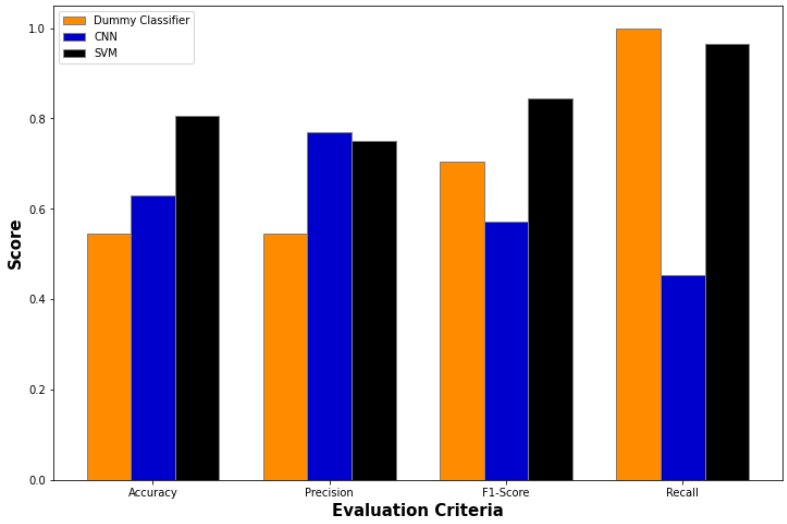
The accuracy, F1 score, precision, and recall parameters would be compared for a baseline, SVM, and CNN as these are the best parameters to evaluate the classification problems. Alongside these, Root Mean Square Error and Mean Absolute Error could also be there in the context, but they are better suited for regression, not classification.

Taking all four parameters is to have a different perspective in the comparison like the F1 score is independent of the true negatives whereas accuracy is not. Same for the precision and the recall.

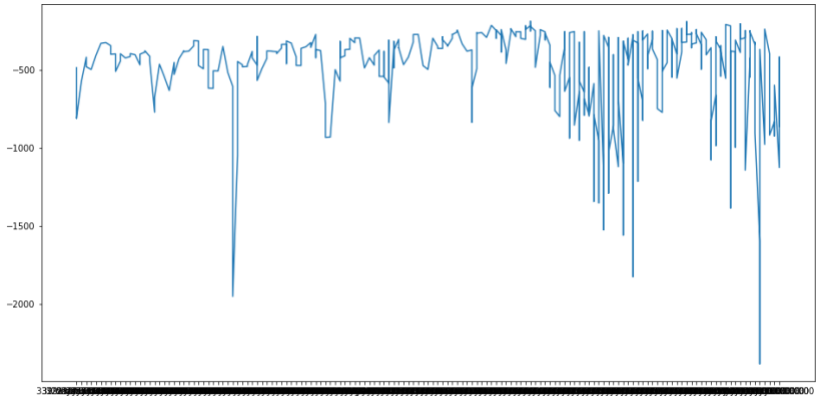
Q-Learning, Q-values vs. Time would be the graph on which the evaluation performed. For Reinforcement learning algorithms, it is not possible to compare it with the preceding models like SVM or CNN.

## PERFORMANCE EVALUATION AND RESULTS

Comparison between Accuracy, Precision, F1-Score and Recall of Dummy Classifier (Baseline model), CNN and SVM model. In the below-plotted chart, we can see a significant improvement on all the parameters compared to the dummy classifier for CNN and SVM.




For the Q table, the traditional metrics would not work as that is reinforming learning algorithm; hence there is a graphical representation of Q-Table values vs. Days to departure, which shows how this benefits the users in cutting the cost down and getting the best deal.



## CONCLUSION AND FUTURE SCOPE

After going through the results of SVM and CNN, it is evident that SVM is slightly better in this context of predicting the correct answer whether the buyer should wait or buy the ticket at the market price right now. Extending ahead, Q-Learning is showing excellent results in the direction of better answer of the agent by looking at the Q-values, the values are improving by a margin, and the plot of Q-values vs. days shows that with higher data points and single targeted model, Q-Learning model would beat both SVM and CNN.

The data used in the model development was collected from the live API, and because of the time constraints, we were only able to fetch the data for ten successive sessions. The number of data points referred to in the project is not sufficient, and the API needs to be triggered for at least 100 sessions for better accuracy and better rewards reflection in Q-Tables.

Future work includes more work in Q-Learning to get better insights into the trend and the rewards, along with the development of DQN and DDQN models. These models with more data points could potentially get the breakthrough.

## REFERENCES

[1] Tianyi Wang, S. Pouyanfar, Haiman Tian, and Shu-Ching Chen, “A Framework for Airfare Price Prediction: A Machine Learning Approach,” *ResearchGate*, Jul-2019. [Online]. Available: https://www.researchgate.net/publication/335936877_A_Framework_for_Airfare_Price_Prediction_A_Machine_Learning_Approach. [Accessed: 03-Dec-2021]

[2] E. J. Santana, S. M. Mastelini, and S. Barbon, “Deep Regressor Stacking for Air Ticket Prices Prediction,” *undefined*, 2017. [Online]. Available: https://www.semanticscholar.org/paper/Deep-Regressor-Stacking-for-Air-Ticket-Prices-Santana-Mastelini/da30a5e2cd62031ddef4a3f75b8583fd8c65d327. [Accessed: 03-Dec-2021]

[3] V. Nair and G. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines” [Online]. Available: https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf

[4] K. Tziridis, Th. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, “Airfare prices prediction using machine learning techniques,” *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug. 2017, DOI: 10.23919/eusipco.2017.8081365. [Online]. Available: https://ieeexplore.ieee.org/document/8081365. [Accessed: 03-Dec-2021]

[5] Vankiet, "A Linear Quantile Mixed Regression Model for Prediction. Bachelor thesis Computer Science Radboud," *document.in*, Sep-2018. [Online]. Available: https://vdocument.in/a-linear-quantile-mixed-regression-model-for-prediction-bachelor-thesis-computer.html. [Accessed: 03-Dec-2021]

[6] R. Ren, Y. Yang, and S. Yuan, “Prediction of Airline Ticket Price” [Online]. Available: https://cs229.stanford.edu/proj2015/211_report.pdf

[7] J. A. Abdella, N. Zaki, K. Shuaib, and F. Khan, “Airline ticket price and demand prediction: A survey,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 375–391, May 2021, doi: 10.1016/j.jksuci.2019.02.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S131915781830884X. [Accessed: 03-Dec-2021]

[8] A. Lantseva, K. Mukhina, A. Nikishova, S. Ivanov, and K. Knyazkov, “Data-driven Modeling of Airlines Pricing,” *Procedia Computer Science*, vol. 66, pp. 267–276, 2015, DOI: 10.1016/j.procs.2015.11.032. [Online]. Available:

https://www.sciencedirect.com/science/article/pii/S1877050915033815. [Accessed: 03-Dec-2021]

[9] B. An, H. Chen, N. Park, and V. S. Subrahmanian, "Data-Driven Frequency-Based Airline Profit Maximization," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2017.

[10] W. Groves and M. Gini, "On Optimizing Airline Ticket Purchase Timing," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 1, pp. 1–28, Oct. 2015, DOI: 10.1145/2733384.

[11] J. Santos Domínguez-Menchero, J. Rivera, and E. Torres-Manzanera, "Optimal purchase timing in the airline market," *ResearchGate*, Aug-2014. [Online]. Available: https://www.researchgate.net/publication/264161565_Optimal_purchase_timing_in_the_airline_market. [Accessed: 03-Dec-2021]

[12] S. Rajankar, N. Sakharkar, and O. Rajankar, "Predicting the Price Of A Flight Ticket With The Use Of Machine Learning Algorithms," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, Dec. 2019 [Online]. Available: http://www.ijstr.org/final-print/dec2019/Predicting-The-Price-Of-A-Flight-Ticket-With-The-Use-Of-Machine-Learning-Algorithms.pdf. [Accessed: 03-Dec-2021]

[13] J. Lu, "Machine learning modeling for time series problem: Predicting flight ticket prices," Feb. 2018 [Online]. Available: https://arxiv.org/pdf/1705.07205.pdf. [Accessed: 03-Dec-2021]

[14] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, May 1992, DOI: 10.1007/bf00992698. [Online]. Available: https://link.springer.com/article/10.1007/BF00992698. [Accessed: 03-Dec-2021]

[15] "Lazy Prediction Library | Flight Price Prediction using Lazy Prediction," *Analytics Vidhya*, 20-Jun-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/flight-price-prediction-a-regression-analysis-using-lazy-prediction/. [Accessed: 03-Dec-2021]

[16] Z. Wang, "A Regression-based Scheme for Flight Price Prediction," 2020 [Online]. Available: https://dalspace.library.dal.ca/bitstream/handle/10222/79128/Wang-Zhenbang-MCSc-CSCI-April-2020.pdf?sequence=1. [Accessed: 03-Dec-2021]

[17] "Bar Plot in Matplotlib," *GeeksforGeeks*, 27-Mar-2020. [Online]. Available: https://www.geeksforgeeks.org/bar-plot-in-matplotlib/. [Accessed: 05-Dec-2021]

[18] C. Zoltan, “SVM and Kernel SVM,” *Medium*, 23-Sep-2021. [Online]. Available: https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200#tuning

[19] “3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.24.1 documentation,” *scikit-learn.org*. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

[20] J. Brownlee, “Evaluate the Performance Of Deep Learning Models in Keras,” *Machine Learning Mastery*, 25-May-2016. [Online]. Available: https://machinelearningmastery.com/evaluate-performance-deep-learning-models-keras/

[21] “1.4. Support Vector Machines — scikit-learn 0.20.3 documentation,” *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

[22] https://www.facebook.com/jason.brownlee.39, “Time Series Forecasting as Supervised Learning,” *Machine Learning Mastery*, 04-Dec-2016. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

[23] “Flowchart Maker & Online Diagram Software,” *app.diagrams.net*. [Online]. Available: https://app.diagrams.net

[24] “Q-Learning in Python - GeeksforGeeks,” *GeeksforGeeks*, 06-Feb-2019. [Online]. Available: https://www.geeksforgeeks.org/q-learning-in-python/

# APPENDIX A: TEAM CONTRIBUTION

| COMPONENT | TEAM MEMBER |
|---|---|
| Data Fetching<br>- API fetching (Kishan, Pathik)<br>- Script for data access (Kishan, Vishal)<br>- Running on the server interval of 48 hours (Dhruv, Vishal) | Dhruv, Kishan, Vishal, Pathik |
| Data Preprocessing<br>- CSV from JSON (Kishan)<br>- Data Formatting (Pathik)<br>- Data Cleaning (Dhruv)<br>- Visualization (Vishal)<br>- Dimension Reduction (Kishan)<br>- One Hot Encoding (Pathik) | Dhruv, Kishan, Vishal, Pathik |
| Models<br>- Baseline (Dhruv, Pathik)<br>- CNN (Dhruv, Pathik)<br>- SVM (Kishan, Vishal)<br>- Q-Learning (Kishan, Vishal) | Dhruv, Kishan, Vishal, Pathik |
| Evaluation<br>- Baseline (Dhruv, Pathik)<br>- CNN (Dhruv, Pathik)<br>- SVM (Kishan, Vishal)<br>- Q-Learning (Kishan, Vishal) | Dhruv, Kishan, Vishal, Pathik |
| Documentation<br>- Research (Dhruv, Kishan)<br>- Diagrams (Vishal)<br>- Document (Pathik, Dhruv) | Dhruv, Kishan, Vishal, Pathik |

## APPENDIX B: CONTRIBUTION PERCENTAGE

| TEAM MEMBER | CONTRIBUTION |
|---|---|
| DHRUV PINTUBHAI DOSHI | 25% |
| VISHAL RAKESH JAISWAL | 25% |
| PATHIK KUMAR PATEL | 25% |
| KISHAN RAKESHBHAI PATEL | 25% |