**Patent form**

1. **Name of the Faculty Mentor:**

Prof. Atul Mishra

1. **Particulars**

|  |  |  |
| --- | --- | --- |
| S No. | Roll No. | Name |
| 1 | 1700247C203 | Nikita Agarwala |
| 2 | 1700218C203 | Surya Kanduri |
| 3 | 1700250C203 | Nishant Pandey |
| 4 | 1700117C203 | Aishwarya Dubey |

2. **Provide a brief descriptive title of your project:**

Flight fare prediction using machine learning

3. **In 50 words or less, please provide an abstract or summary of your project:**

Aircrafts use different computational models to expand their income, for example, request forecast and value separation. In our project we try anticipating the base price for the consumers for various routes in regards with the Indian market.We see how the base price depends upon the services being provided as well as the number of stops that are being taken.We show that by stacking of ML models and doing ensemble learning we are able to handle this regression problem with 92% accuracy after extensive data preprocessing.

4. **State the problem or problems that motivated or required a solution provided by this project:**

The project deals with airfare prices prediction. Airlines use various algorithms to fluctuate and control the airfares in order to increase their revenue. According to Google Trends, "Cheap Air Tickets" is the most searched query in the aviation sector in India. We will focus on aspects that are visible on the consumer side and try to predict the base fare for a given route. Also we have included the price for the tickets on the number of stops and weekends, so this will help customers to choose the flight ticket effectively.

5. **List the specific problem which your project is solving:**

From the client side, two sorts of models are proposed by various analysts to set aside cash for clients: models that foresee the ideal time to purchase a ticket and models that anticipate the base ticket cost. In our project we anticipate the base price for the consumers for various routes in regards with the Indian market..We model various ML algorithms to do the above mentioned work and then compare their performance

**6. Provide a detailed explanation of how this project solves the problem(s).**

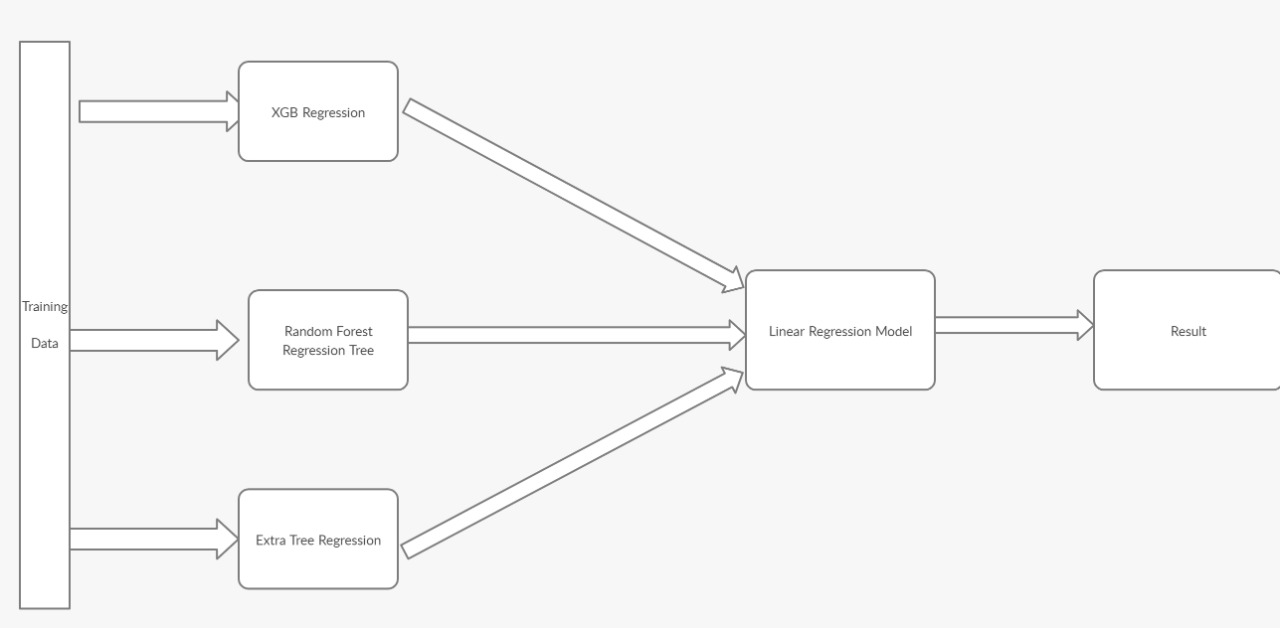
We tackled our problem in phases.The phases are discussed below in detail:

1. **Data Collection-** Collection of data using APIs from Expedia and various other sites.Manually collecting data daily is not efficient and thus a python script was run on a remote server which collected prices daily at specific time .Our database was of the shape [10463,11].10 features are intrinsic in nature and ‘Price’ is the label for our project work.
2. **Cleaning Of Data/Data Preprocessing-** On analysing the data we found that there were many empty rows and columns with null value that were of no use so we dropped them.
   1. Replace repeating values in the data set :For example,in the database New Delhi and Delhi both were used.So we changed all values of Delhi into New Delhi.
   2. Converting ‘Duration of Flight’ and ‘Date’ into workable format.
3. **Feature Selection-** During this phase we have identified various features which are required for accurate prediction of flight prices. For every flight we have collected the features like DOJ,Airline Name,Source,Destination,Route,Departure Time,Arrival Time,Total Stops,Holiday (yes/no).We can also do feature generation and find important features from each model to build our final model.
4. **Encoding the data :**Since the regression models work better with numerical inputs we encoded features like ’Source’,’Destination’,’Additional\_Info’,’Day’ and different routes using Label Encoder and ‘Airlines’ using One-Hot Encoder.Label encoder converts categorical data into numeric values but this can misinterpreted by algorithms as ordinal data.So we use ‘One-Hot Encoder’ of scikit.It does binary encoding and adds a new column in the data frame for each unique category value.So for “Airlines’ we got 12 new columns and that is the disadvantage.It swells the dataset thus the execution time also increase.So we can do Principal Component Analysis for Dimensionality Reduction of our dataset.
5. **Attribute construction**:We constructed some new features out of the existing features to get more information.For example we further broke Route Attribute into 5 different features as it was important.A value like BLR->DEL was broken individually into BLR and DEL as intuitively path taken has some effect on airfare.
6. **Model Building-** In our project we have used following 4 models:

|  |  |
| --- | --- |
| *ML Model* | *R squared after One-Hot Encoder+After normalization* |
| Linear Regression | 0.60 |
| Random Forest Regression Tree | **0.90** |
| Extremely Randomized Trees | 0.90 |
| XGB Regressor | 0.89 |
| Ensemble Learning-Stacking | **0.92\*** |

\*Read R2 as 92% of the variation in Price can be explained by the other features

1. **Ensemble Learning :**Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem.Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features.So in our case,we did stacking as shown below.

****

**7. Existing state-of-the-art: (Brief background of the existing knowledge.)**

There are several previous studies regarding flight price prediction using machine learning algorithms.

* K. Tziridis, through his research paper “Airfare Prices Prediction, Using Machine Learning Techniques” has shown that it is feasible to use ML models for predicting airfare prices.He has claimed that “Bagging Regression Tree”, “Random Forest Regression Tree”, “Regression Tree” are the most stable model with variation in their accuracy scores.He has build all these models again and again with considering different features like departure time and overnight journey (yes/no).If we consider execution time as a parameter then “Random Forest Regression Tree” and “Regression tree” are the best performing. But he has worked on some Greek airlines.
* Vinal Raja and his colleagues have worked on this problem statement in context with the Indian aviation sector and in their model they have tried predicting that whether it is the right time to buy a ticket or not.In their work Logistic Regression,Regression Tree and Bagging Regression Tree have given high accuracy.They also found out that for some routes like Mumbai-Delhi prices remain constant and they attributed this behaviour to high frequency of flights ,heavy competition and high demand.
* Gini and Groves took the Partial Least Square Regression(PLSR) for developing a model of predicting the best purchase time for flight tickets. The data was collected from major travel journey booking websites from 22 February 2011 to 23 June 2011.
* Janssen built up an expectation model utilizing the Linear Quantile Blended Regression strategy for SanFrancisco to NewYork course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the takeoff date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the takeoff date, anyway for a considerable length of time close to the takeoff date, the expectation isn't compelling.
* A study by Dominguez-Menchero recommends the ideal buying time dependent on nonparametric isotonic relapse method for a particular course, carriers, and timeframe. The model gives the most extreme number of days before buying a flight ticket.The passage taken and date of procurement are the two variables that are considered for expectation

**8. List out the known ways about how others have tried to solve the same or similar problems? Indicate the disadvantages of these approaches. In addition, please identify any prior art documentation or other material that explains or provides examples of such prior art efforts.**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Existing state of art** | **Drawbacks in existing state of art** | **Overcome** |
| 1 | Airfare Prices Prediction Using  Machine Learning Techniques by K. Tziridis, Th. Kalampokas, G.A. Papakostas | * It was a preliminary study that showed that it is feasible to predict prices for flights based on historical fare data and they did for a single greek airline | We did it in regards with Indian market and did for many airline.We also add ensemble learning and one hot encoder learning to it |
| **2** | Prediction of Airfare Using Machine Learning by Vinal Raja,Janhavi Vakil,  Yash Shah, Sonia Relan | They have tried predicting whether it is the right time to buy a ticket or not and used different models for it. Their problem was a classification problem. | We used XGB Regression,Random Forest Regression Tree  Extra Tree Regression and showed that with some data preprocessing these models can give higher accuracy |
| **3** | Predicting The Price Of A Flight Ticket With The  Use Of Machine Learning Algorithm by  Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar | They used basic ML models like Decision tree ,Random forest,K-NN and Linear Regression. These models have their own disadvantages.Their accuracy is 0.67 which is low. | We suggest more preprocessing like normalization ,encoding and attribute construction to make our data more useful.We also stacked various ML models which gave us better results. |

**9. List the Technical features and Elements of the project.**

A. XGB Regression

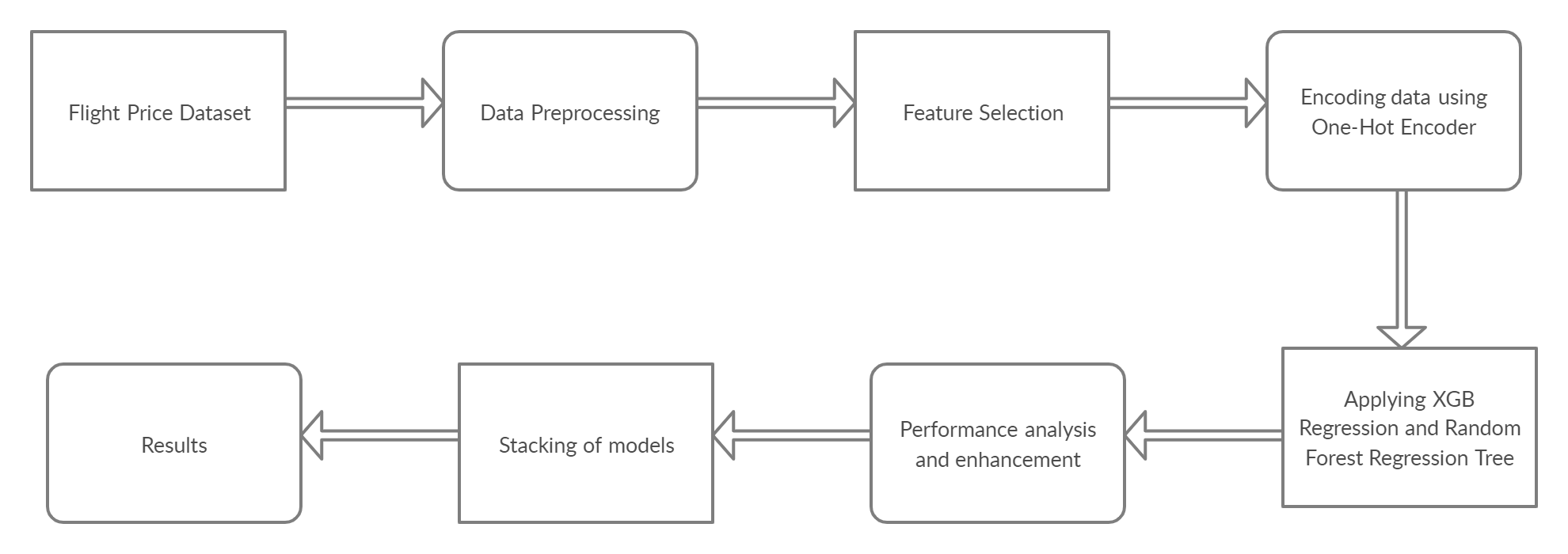
B. Random Forest Regression Tree

C. Extra Tree Regression

D. Linear Regression

E. Ensemble learning-stacking

**10. Draw the block diagram of your project**

****

**11. List all the components (hardware and software used in your project)**

Software Requirements-

1. Windows 10/MacOs Catalina
2. Python 3.4(Google Colab)

**12. List out the features of your project which are believed to be new and distinguish them over the closest technology.**

* Emphasis on Use of One Hot Encoder instead of Label encoder for this problem.It is better to use One Hot Encoder on features like ‘Airlines’,’Source’,’Destination’ and ‘Additional Info’ as these features are categorical in nature but not ordinal .So intuitively we should not use label encoder for encoding such features. and rather use One Hot Encoder.It will be wrong to use label encoder because it introduces ordering in our dataset we should be avoided for problem like these.For example, it is not right to encode “Jet Airways” as 1 and Air India as “2” as this introduces ordering and that affects our results by overfitting the data.
* The use of One Hot Encoder will swell up any dataset so we suggest using Principal Component Analysis for reducing the dimensions in our dataset before training ML models for this problem.
* Also,we suggest the use of Ensemble Learning in the problem of predicting airfare as it helps us to overcome any unfair bias that is introduced because of one single model.We did stacking and it gave us accuracy of around 90-92% and that is fair because our dataset did not have distance as a feature and distance does have an impact on the airfare.

**13. Are there alternative ways of implementing your project that is different from what you have disclosed? Specifically, if someone knew of your solution to the problem you solved (Question 3), would it be easy for them to come up with an alternative solution to the same problem that did not include details of your project?**

**Please explain:**

In Datasets like these we need to do a lot of data preprocessing and feature engineering and everyone can have their own approach for the same.There are some similar approaches to this problem and anyone can improvise our idea with the same or a different algorithm.But our claim that binary encoding in flight datasets and using ensemble learning will help everyone in a long run.

**14. Status of your project: been built or tested or implemented? If so, please provide the particulars of the first time it was successfully built or implemented (when, where, by whom, and evidence of this event including written or on-line pointers to documentary evidence):**

Due to unavoidable circumstances due to Covid-19 we could not completely execute the automatic script written to mine data automatically so we tested for a certain dataset available on MachineHack.

The link to the database is:

<https://www.machinehack.com/course/predict-the-flight-ticket-price-hackathon/>

The dataset was applied to our model and certainresults have been observed which have the accuracy measure of around 90% after stacking various ML models.

GitHub link to the tested project:

<https://github.com/nikitaagarwala16/Predicting-Airfare-Price>

**15. Briefly state when and how you first conceived this idea?**

As a course requirement for Machine Learning and Data Mining we were supposed to work on a project .As we all live far from our college and we often take flight back home so we thought why not work on a project where we could predict the base fare for our flight tickets and so we choose this project.

**16. Please provide the names of products that your project will be used in (if any):**

None

**.17. Additional Information:**

This is a very data extensive problem.More the number of data points better would be our results.Also we cannot have a 100% accuracy in this problem because the pricing of air tickets are affected by a number of hidden factors which are only known by these airlines giants but we can use machine learning to have a good approximate about airfare.

**18. References:**

1. [Airfare Prices Prediction Using Machine Learning Techniques by K. Tziridis, Th. Kalampokas, G.A. Papakostas](https://ieeexplore.ieee.org/document/8081365)
2. [Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithm by Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar](http://www.ijstr.org/final-print/dec2019/Predicting-The-Price-Of-A-Flight-Ticket-With-The-Use-Of-Machine-Learning-Algorithms.pdf)
3. [Prediction of Airfare Using Machine Learning by Vinal Raja,Janhavi Vakil,](http://www.ijsdr.org/papers/IJSDR1804041.pdf)

[Yash Shah, Sonia Relan](http://www.ijsdr.org/papers/IJSDR1804041.pdf)