**Proof of Concept (Flight Fare Price Prediction)**

**Flight Fare Price Prediction**

**Problem Statement**

The problem statement is based on certain independent variables like Airline, Date of Journey, Source, Destination, Route, Arrival time and so on. Price of flight will be predicted based on these factors

**Objective**

Develop a predictive model for predicting the price of flight based on available data

**Benefits**

* Helps us to easily estimate the flight price across the length and breadth of India

**Source of Training and Test Data**

The dataset was generated from Kaggle and it contains 10683 training data. Some of the attributes of the training data include Airline, Date of Journey, Source, Destination, Route, Departure Time, Arrival Time, Duration, Total Stop, Additional Information and Price

**Data Preprocessing**

* Price which is the dependent variable is the only feature with integer data type, all others are object data type. All other columns were converted to numeric
* Using the pandas date time to convert object data type to date time data type
* Feature engineering on some of the features of the data
* Checking for outliers
* Standardizing the data

**Model Selection and Training**

The training data is fit on five algorithms, "Linear regression", "Decision Tree" "SVM", "Random Forest" and "XGBoost". XGBoost performed the best and a pickle file was created where the model was saved for the purpose of reusability

**Prediction Data Description**

**The test data contains 2671 data with all features of the training data except price**

**Data Preprocessing**

All the data preprocessing technique that was applied on the training data will also be replicated on the prediction data for the model performance too be okay

**Deployment**

The model was deployed on Amazon Web Service (AWS) and platform as a service

**Question and Answers**

**Q1. Tell me about your project.**

Here is a project that gives us a clear idea of what the flight price is moving across different parts of India based on certain features of the data such as Airline, Date of Journey, Source, Destination, Route, Departure Time, Arrival Time, Duration, Total Stop and Additional Information

**Q2. What is the source of the data?**

The data for training is provided was gotten from Kaggle

**Q3. What are the techniques used for data preprocessing?**

* Removing unwanted attributes
* Visualizing the relationship between independent variables with each other and output variables
* Removing columns with zero standard deviation
* Cleaning data and imputing of null values where necessary
* Converting categorical data into numeric values
* Removal of outliers
* Scaling the data

**Q7. How training was done or what model was used?**

Single model was not used on the entire training data. Algorithms like SVM, XGBoost, and Random Forest were used and based on accuracy score, a final pickle file was created where the model with the highest accuracy was stored

**Q8. How was prediction done?**

The testing file from Kaggle was also used for the prediction .We performed the same life cycle that was performed on the testing data for the prediction to be easily made. In the end we get the accumulated data of predictions

**Q9. What was the data type?**

The data used in the training of this model is a combination of both numerical and categorical values

**Q10. What was the team size and distribution?**

This was just a POC, I basically performed all the work

**Q11. What is the kind of automation for data processing?**

A full-fledged ETL pipeline is in place for data extraction processing and loading into the database

**Q12. Have you used any scheduler?**

NO! A scheduler was not used to retrain the model

**Q13. How are you monitoring job?**

There are logging set-ups done. We regularly monitor the logs to see for any error scenarios. For fatal errors, we had email notifications in place. Whenever a specific error code, which has been classified as a fatal error occurs, email gets triggered to the concerned parties.

**Q14. What were your roles and responsibilities in the project?**

I participated in all the activities to make the project work. Both core and non-core data science activities

**Q15. In which area you have contributed the most?**

My contribution was the reason the project was completed. I participated in all

**Q16.In which technology you are most comfortable?**

I have worked across diverse areas in the industry using different technological stack ranging from Machine Learning, Deep Learning and NLP. With my experience and exposure, I love working in Machine Learning and Deep Learning

**Q17. How you rate yourself in big data technology?**

Being someone that is hungry for growth across the tech industry, Big data is the area I am working seriously on to become proficient. I’m currently learning SPARK

**Q18. In how many projects you have already worked?**

Giving a precise figure will be difficult but I have worked in various small and large scale projects e.g Deep learning, Computer Vision, NLP projects, Chatbot building, Machine learning with regression, and classification problems.

**Q19. How would you rate yourself in distributed computation?**

I will rate myself 7

**Q20. In which part of machine learning have you already worked on?**

I have worked on both supervised and unsupervised machine learning approaches and building different models using the as per the requirement of the client/user

**Q21. What are the areas of machine learning algorithms that you already have explored?**

I have explored various machine learning algorithms like Linear Regression, Logistic Regression, L1 and L2 Regression, Polynomial Regression, Multi Linear Regression, Decision Trees, Random Forests, Extra Trees Classifier, PCA, XGBoost, CAT Boost, ADA Boost, Gradient Boosting, Light Boost, K-Means, K-Means ++, LDA, QDA, KNN, SVM, SVR, Naïve Bayes, Agglomerative clustering, DBSCAN, Hierarchical clustering and so on

**Q22. How much time did your model take to get trained?**

With a training batch size of 41 and a laptop with system configuration of 64GB RAM using NVIDIA Pascal Titan GPU, the entire training took 45 minutes

**Q23. At what frequency are you retraining and updating your model?**

The model gets retrained every 20 days.

**Q24. How would you rate yourself in machine learning?**

On the scale of 10, I will rate myself 8

**Q25. Why are you leaving your current organization?**

Being in the organization was a blessing but what we do is becoming monotonous and boring. The quest to explore and seek for knowledge, skills and a fresh challenge is the reason behind my job search

**Q26. In which mode have you deployed your model?**

I have deployed the model both in cloud environments as well in the on-premise ones based on the client and project requirements

**Q27. How were you doing deployment?**

The mechanism of deployment depends on the client's requirement. I deployed this project on two cloud platforms which are AWS and Platform as a service