Practical - 01.

- **Aim:** Predict the price of the Uber ride from a given pickup point to the agreed drop off location.
  Perform following tasks -
  1. Pre-process the dataset.
  2. Identify outliers.
  3. Check the correlation.
  4. Implement linear regression and random forest regression models.
  5. Evaluate the models and compare their respective scores like R2, RMSE, etc.

- **Theory** :

1) Dataset Preprocessing -

- It is a crucial step to ensure the dataset is clean, consistent and suitable for modeling. The uber fares dataset contains features such as pickup and drop-off longitude, latitude, timestamp, passenger count, and fare amount.
- Handling missing values - Null or empty rows must be removed to avoid biased training.

- Feature extraction - From the date time column, additional features like hour of the day, day of the week and month can be extracted to capture temporal effects on fares.

- Data type conversion – Ensuring proper formats.
- Feature scaling – Normalization, or standardization may be applied to numerical features if required for regression models.

2] OUTLIER DETECTION –

- Outliers can heavily skew regression models and result in poor predictions.
- Hence outliers such as negative fares, unrealistic passenger counts or invalid co-ordinates are removed for cleaner model training.

3] Linear regression – [In depth] –

- It is a supervised machine learning algorithm that models the relationship between independent variables (features) and the dependent variable (fare amount) by fitting a straight line [or hyperplane in multiple dimensions.

Mathematical formulation –

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Page No. 3
Date

Where, y = predicted fare amount.
$x_1, x_2 \dots x_n$ = independent variables (distance, time, etc).
β_0 = intercept.
β_1 = co-efficients for each feature.
ϵ = error term.

Key characteristics of linear regression are -

1. Interpretability - Each coefficient explains how much the fare changes with a unit change in that feature (eg. higher distance directly increases the fare.)

2. Assumptions - Assumes linearity, independence of errors, equal variance of residuals, and normally distributed errors.

3. Strengths - Simple, fast to train, provides interpretable results.

4. Limitations - Cannot capture non linear patterns in fares, (eg. surgepricing, geographical constraints)
- Highly sensitive to outliers.

It works well as a baseline but often underfits real world datasets like uber fares where pricing depends on complex and non linear relationships.

4] Random Forest Regression -

Page No.
Date

Random forest is an ensemble learning algorithm based on decision trees. It builds multiple regression trees on random subsets of data and averages their predictions to make the final output.

How it works –

1. Bootstrap Sampling – Creates multiple random samples from the training data (with replacement).

2. Decision Trees – For each example a regression tree is trained. At each split, a random subset of features is considered, which introduces diversity among trees.

3. Aggregation – predictions from all trees are averaged to produce the final fare prediction.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} f_i(x)$$

Where,

k = number of trees in the forest.

$f_i(x)$ = prediction from i^{th} tree.

* Characteristics –

- Handles Nonlinearity - Can model complex interactions between features (eg. distance + rush hour + location)

- Robust to outliers and noise - Outliers in individual trees have less effect when results are averaged.

- Feature importance - Provides measures of which features (distance, time, passenger, count) influence fares the most.

- Hyper parameters - Includes number of trees (n_estimators), tree depth (max_depth) and minimum samples per split (min_samples_split).

- Strengths -
  - High predictive accuracy.
  - Can generalize well without heavy assumptions (no need for linearity)
  - Works well with large datasets.

- Limitations -
  - Less interpretable compared to linear regression.
  - computationally more expensive (training multiple trees)
  - May overfit if the forest is too deep or not regularized properly.

In practice, random forest regression often outperforms linear regression in Uber fare prediction due to the non linear and context dependent nature of pricing.

5) Model Evaluation –

1. R^2 score – Measures proportion of values variance in fares explained by the model.

2. RMSE – Root Mean Squared error – Penalizes larger errors more.

3. MAE – Mean absolute error – Average absolute prediction error.

Expected outcome –

- Linear regression – Moderate R^2, higher RMSE.
- Random Forest : Higher R^2, lower RMSE and MAE making it more accurate.

6] CONCLUSION:

The practical successfully demonstrated the application of regression models to predict Uber rides fares. While linear regression provided an interpretable baseline, it was limited by assumptions of linearity and sensitivity to outliers. Random forest regression by contrast acheived a superiour predictive performance. Thus, proving to be an effective model.