Capstone Final Project Report

Title: *Capstone Two – Taxi Trip Price Prediction*

1. Problem Identification

Taxi companies need accurate pricing strategies that reflect trip distance, duration, traffic, and weather conditions. Inefficient pricing can either lose revenue (if underpriced) or reduce customer satisfaction (if overpriced).
The goal of this project is to **predict taxi trip prices** using regression models. This will help improve pricing strategies and enhance customer trust.

2. Dataset

- **Target Variable:** `Trip_Price` (continuous)
- **Features:**
  - Numeric: Trip_Distance_km, Passenger_Count, Base_Fare, Per_Km_Rate, Per_Minute_Rate, Trip_Duration_Minutes
  - Categorical: Time_of_Day, Day_of_Week, Traffic_Conditions, Weather

Data Types:

- Float: numeric features
- Object: categorical features

3. Data Preparation

- **Missing Values:** Imputed using SimpleImputer with mean strategy.
- **Categorical Encoding:** Applied one-hot encoding (pd.get_dummies) to categorical variables.
- **Feature Scaling:** Standardized numeric variables with StandardScaler.
- **Train-Test Split:** 80% train, 20% test.

4. Exploratory Data Analysis (EDA)

- Scatterplot showed a **positive correlation** between Trip_Distance_km and Trip_Price.

- Boxplot analysis revealed **higher trip prices** under heavy traffic and adverse weather.
- Distribution plots indicated that trip duration and distance were skewed, impacting pricing.

Figures Included:

1. Scatterplot: Trip Distance vs Price
2. Boxplot: Price by Traffic Conditions
3. Distribution Plot: Trip Price

5. Models Built

Three regression models were trained and tuned using **GridSearchCV**:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor

6. Model Performance Comparison

Model	MAE	MSE	R^2
Linear Regression	**9.835**	**193.902**	**0.7665**
Random Forest	**5.401**	**59.848**	**0.9279**
Gradient Boosting	**4.954**	**56.612**	**0.9318**

Best Model: Gradient Boosting Regressor

7. Final Model Application

The Gradient Boosting Regressor was selected as the final model due to its superior performance (lowest error rates, highest R^2 score). It was applied to the test set, producing accurate predictions of taxi fares.

8. Recommendations

1. Deploy Gradient Boosting Model into production for fare estimation.
2. Incorporate real-time traffic and weather data to further improve predictions.
3. Use feature importance analysis to refine base fares and surcharges for specific conditions.

9. Future Work

- Test advanced ensemble models (e.g., XGBoost, LightGBM).
- Incorporate external data (special events, holidays, surge demand).
- Develop a real-time prediction system integrated with GPS tracking.

---

10. Conclusion

This project successfully demonstrated how machine learning models can be applied to predict taxi trip prices. After data wrangling, EDA, and model comparison, the Gradient Boosting Regressor was identified as the best-performing model. Its deployment can help taxi companies improve pricing accuracy, customer satisfaction, and profitability.