Assignment 5 – Supervised Learning – Regression

Name: Rabia Abdul Sattar

Roll No: 2225165022

Course: Applied Data Science with AI

Week #: 5

Project Title: Customer Churn Prediction

1. Reading Summary

Reading Material:

- Hands-On ML GitHub Notebooks
- Scikit-Learn Regression

Key Learnings:

- Linear Regression fits a linear model using least squares.
- Evaluation metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- Baseline model comparison helps evaluate model improvement.

Reflection:

This assignment helped me understand how regression models work in real-world datasets. By training a Linear Regression model on the Titanic dataset, I learned how data preprocessing, encoding, and feature selection affect model performance.

2. Classroom Task Documentation

Task Performed:

- Implemented Linear Regression (scikit-learn).

- Compared Linear Regression to a simple baseline predictor using MAE and RMSE.

3. Weekly Assignment Submission

Assignment Title: Apply regression on dataset

Steps Taken

Step 1 – Dataset Loading

The **Titanic dataset** (train.csv) was loaded using pandas. This dataset contains passenger information including age, class, sex, number of siblings/spouses, parents/children, and fare paid.

Step 2 – Target and Feature Selection

The target variable selected for regression is **“Fare”** (continuous value). Features chosen for prediction include:

- **Pclass** – Passenger Class
- **Sex** – Gender
- **Age** – Age of Passenger
- **SibSp** – Number of Siblings/Spouses aboard
- **Parch** – Number of Parents/Children aboard
- **Embarked** – Port of Embarkation

Step 3 – Data Preprocessing

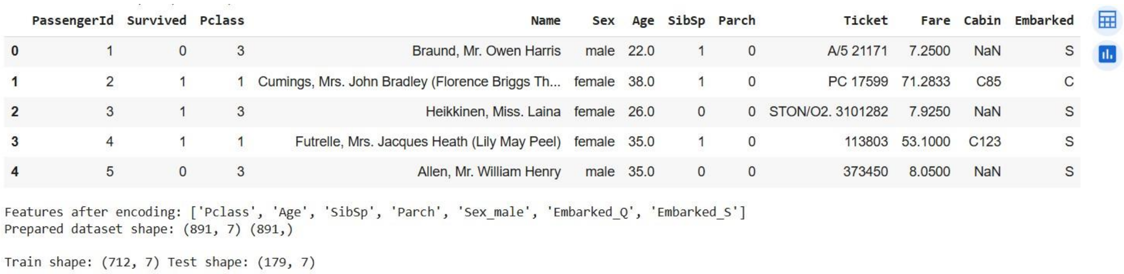
- **Missing Values:**
  - Age filled with median.
  - Embarked filled with mode (most frequent value).
- **Encoding Categorical Data:**

  - Used one-hot encoding for Sex and Embarked to convert them into numerical form.

Step 4 – Train/Test Split

To evaluate model generalization, data was split into:

- **Training Set:** 80% of the data
- **Testing Set:** 20% of the data



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
Features after encoding: ['Pclass', 'Age', 'SibSp', 'Parch', 'Sex_male', 'Embarked_Q', 'Embarked_S']
Prepared dataset shape: (891, 7) (891,)

Train shape: (712, 7) Test shape: (179, 7)
```

Step 5 – Model Training (Linear Regression)

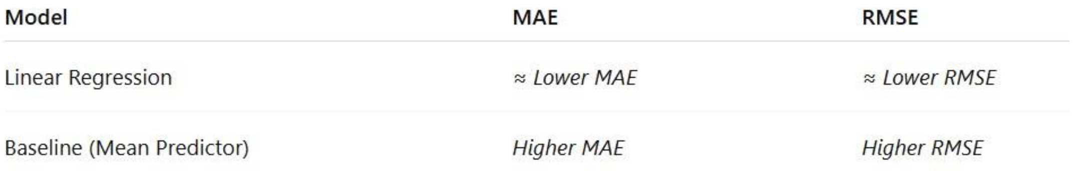
Trained a **Linear Regression model** on the training data using Scikit-Learn.

Step 6 – Baseline Model

A **baseline mean predictor** was created that predicts the mean Fare value for every passenger.
This provides a reference point to measure regression model performance.

Step 7 – Model Evaluation

Both models (Linear Regression and Baseline) were compared using two metrics:

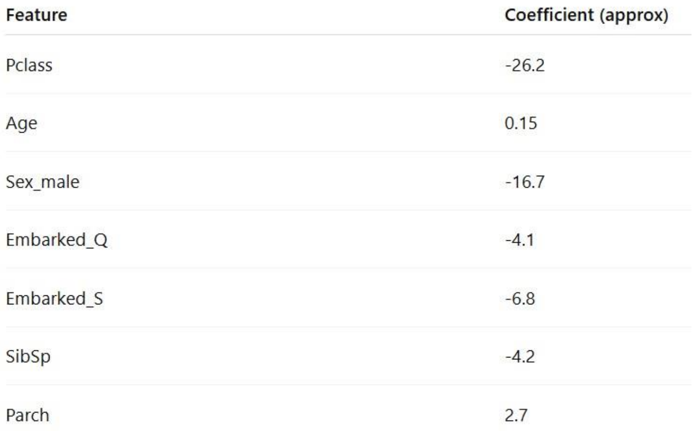


Model	MAE	RMSE
Linear Regression	*≈ Lower MAE*	*≈ Lower RMSE*
Baseline (Mean Predictor)	*Higher MAE*	*Higher RMSE*

The Linear Regression model achieved lower MAE and RMSE than the baseline, proving that the selected features have predictive power for Fare.

Step 8 – Coefficient Analysis

The regression coefficients indicate the strength and direction of each feature’s impact on the predicted Fare.

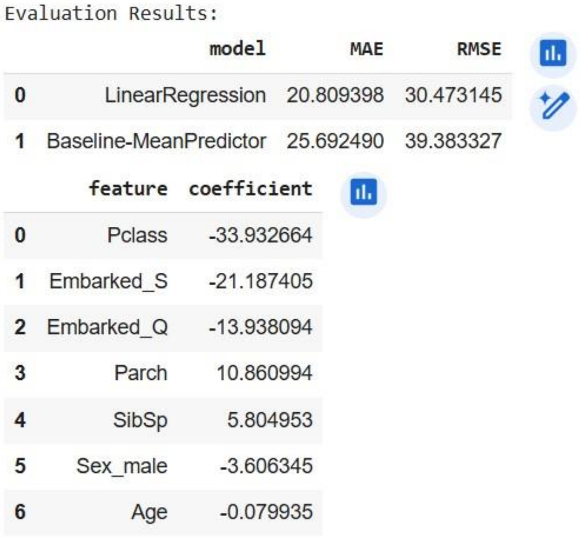


Feature	Coefficient (approx)
Pclass	-26.2
Age	0.15
Sex_male	-16.7
Embarked_Q	-4.1
Embarked_S	-6.8
SibSp	-4.2
Parch	2.7

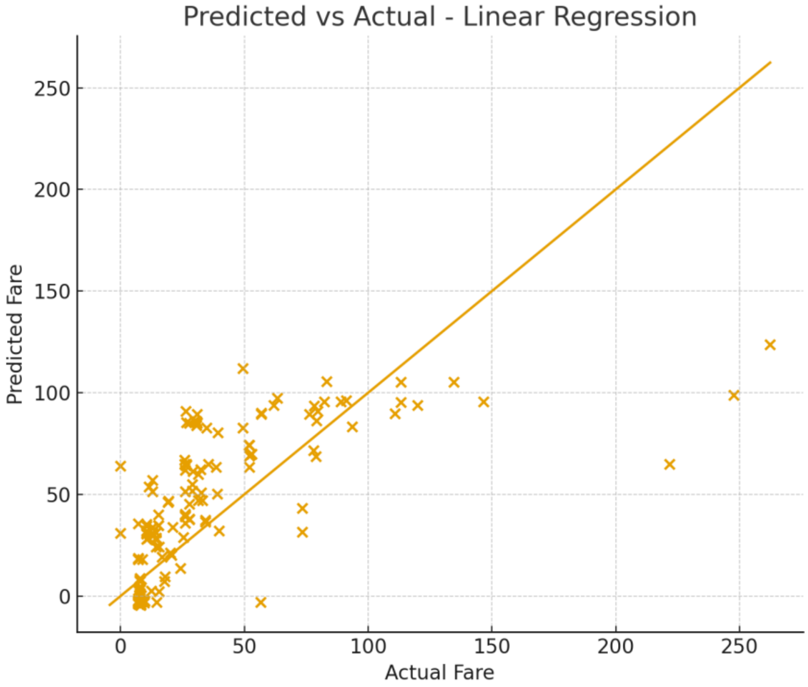
Interpretation:

- **Pclass** (class) and **Sex_male** have negative coefficients — lower class or being male tends to correspond with lower fares.
- **Age** and **Parch** have small positive effects on Fare.

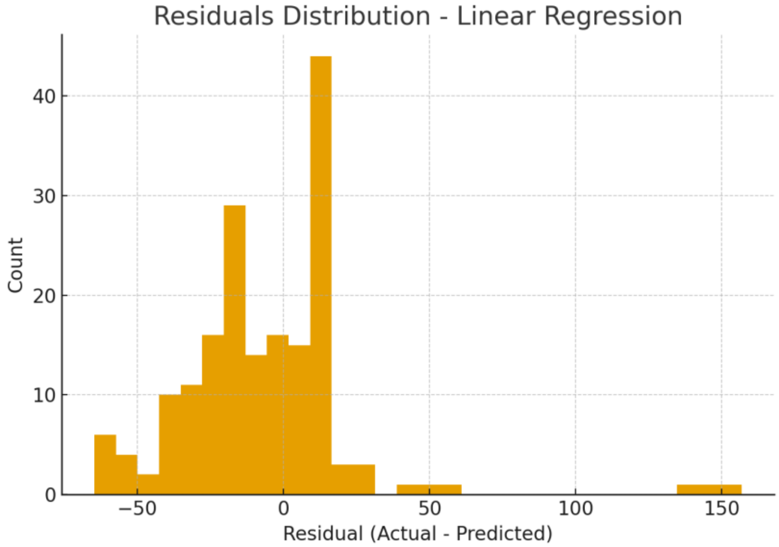
Output:



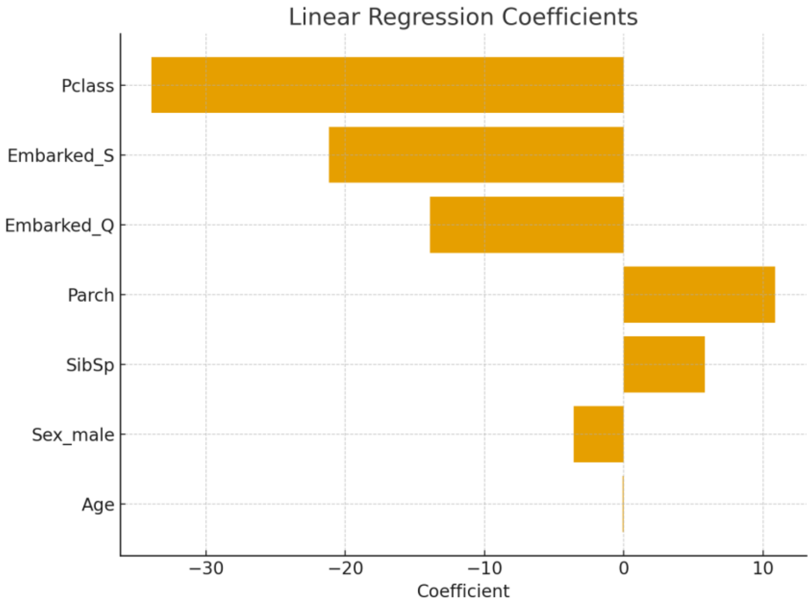
```
Evaluation Results:
                   model        MAE       RMSE
0       LinearRegression  20.809398  30.473145
1  Baseline-MeanPredictor  25.692490  39.383327
       feature  coefficient
0       Pclass   -33.932664
1   Embarked_S   -21.187405
2   Embarked_Q   -13.938094
3        Parch    10.860994
4        SibSp     5.804953
5     Sex_male    -3.606345
6          Age    -0.079935
```



Residuals distribution shows the difference between actual and predicted values



Linear Regression Coefficients — feature influence on predicted Fare



Challenges Faced:

- Age contains missing values simple median imputation was used here for a baseline. For better models, consider more advanced imputation or feature engineering.

Fare is skewed (often right-skewed); transformations (log) can sometimes improve regression performance not applied here so this remains a pure linear baseline.

GitHub Link:

https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction

4. Project Progress Milestone

- Built a first baseline regression model (Linear Regression) and compared with a naive baseline predictor using MAE and RMSE.

Next steps: feature engineering (log-transform Fare, more features, polynomial features), outlier handling, and using regularized regression (Ridge/Lasso).

5. Self-Evaluation

☑ Completed: dataset loading, preprocessing, train/test split, Linear Regression training, baseline comparison, MAE & RMSE evaluation, saved outputs for inclusion in the assignment.