Report
Assignment-4

Cs19btech11020

Q5:
(i)What is the logistic model P(ˆy = 1|x1, x2) and cross-entropy error Function?

Accuracy=66.66
The logistic model: W= [2.61239626 0.84427269] , b= -1.8091969298585213
Cross-Entropy error=

Epoch: 0 Loss= -0.6931471805599453
Epoch: 200 Loss= -0.6336049880786752
Epoch: 400 Loss= -0.5849779704868059
Epoch: 600 Loss= -0.5422712682948627
Epoch: 800 Loss= -0.5046626269625096
Epoch: 1000 Loss= -0.47143172960981133
Epoch: 1200 Loss= -0.44195701563626977
Epoch: 1400 Loss= -0.41570763419647266
Epoch: 1600 Loss= -0.39223307692049025
Epoch: 1800 Loss= -0.3711523176194232
Epoch: 2000 Loss= -0.35214348834466674
Epoch: 2200 Loss= -0.3349345895567221
Epoch: 2400 Loss= -0.3192954078490455
Epoch: 2600 Loss= -0.30503063336427944
Epoch: 2800 Loss= -0.2919740785866871
Epoch: 3000 Loss= -0.27998386367222033
Epoch: 3200 Loss= -0.26893842682998453
Epoch: 3400 Loss= -0.2587332267479291
Epoch: 3600 Loss= -0.24927801908552047
Epoch: 3800 Loss= -0.24049460592120406
Epoch: 4000 Loss= -0.23231497333347165
Epoch: 4200 Loss= -0.2246797469239634
Epoch: 4400 Loss= -0.2175369076951931
Epoch: 4600 Loss= -0.2108407212821074
Epoch: 4800 Loss= -0.20455084228630893
Epoch: 5000 Loss= -0.1986315626188717

(ii)Use gradient descent to update θ0, θ1, θ2 for one iteration. Write down the updated logistic regression model.

Updated logistic model :
θ0=-1.0031662597725644
θ1=1.50535086
θ2=0.5019686

(iii)At convergence of gradient descent, use the model to make predictions for all the samples in the test dataset. Calculate and report the accuracy, precision and recall to evaluate this model.

Accuracy=66.6
Precision=0.6
Recall=1

Q(6)
Pre-Processing data

- Removing nan values from data.
- Adding date ,time ,week, year, hour, columns by splitting pickup time column for understanding peek times.
- Adding Distance column by calculating the distance using latitude and longitude, as it has high correlation with the target variable
- Removing Outliers which are not present in test set as they wont contribute towards the prediction

Trained only on a subset of 5,00,000 samples.
Lowest RMSE was obtained on XGBoostRegressor:

XgboostRegressor:`(objective='reg:linear', n_estimators=100, max_depth=5, n_jobs=-1, random_state=42)`
RMSE= 3.31504

RandomforestRegressor(`n_estimators=100,max_depth=5,max_features=7,random_state=0)`
RMSE=3.84415

XGBoost always gives more importance to functional space when reducing the cost of a model while Random Forest tries to give more preferences to hyperparameters to optimize the model. Random Forest is a bagging algorithm. It reduces variance. Boosting reduces variance, and also reduces bias. It reduces variance because we are using multiple models (bagging). It reduces bias by training the subsequent model by telling what errors the previous models made (the boosting part).