

Task 3 – Model Design Specification:

Created a feature matrix X

Target Vector y (Interest Rate)

Used these Variable for training and testing

Used LinearRegressor module from Model Selection

Used RandomForestRegressor from Ensemble

Used DNNRegressor from tensorflow

Independent/Dependent Variable

Dependent Variables:

Subgrade:

Subgrade is the extension of the grade. Grade classifies a borrower from in range from A-G

And subgrade gives us detailed classification of borrower grade from A1-A6 to G1-G6.

Funded_amnt:

Funded_amnt is dependent on Loan_amnt .Funded_amnt is always less than or equal to the Loan_amnt.

Independent variables:

Loan_id
Member_id
Loan_amnt
Funded_amnt
Term
Grade
emp-length
home_membership
annual-inc
Loan-Status
purpose
state_add
dti
delinq_2yrs
total_acc

Summarize MAPE (Mean Absolute Percentage Error) for training and testing data

$$\frac{ABS(Actuals - Forecast)}{Actuals} * 100$$

MAPE value for Linear Regression → 3.9

MAPE value for Random Forest → 1.6

MAPE value for Neural Networks(DNN Regressor) → 8.79

From the above information, it is concluded that MAPE for random forest is less so the interest rate predicted using this model will be the best model to go with by using MAPE as a criteria.

5 Fold Cross Validation

```
In [40]: ▶ #k = 5 fold cross validation for Linear Regression
from sklearn.model_selection import cross_val_score
scores = cross_val_score(estimator = regressor,X = X_train, y = y_train, cv=5)
```

```
In [43]: ▶ #Average of all the folds and prediction the accuracy
scores.mean()
```

```
Out[43]: 0.9724744967042387
```

```
In [44]: ▶ #Standard Deviation Calculation
scores.std()
```

```
Out[44]: 0.00017648873901682843
```

Dividing the dataset into folds and test in each folds which improves the performance of the model

- Linear Regression → 97% accuracy
- Random Forest Regressor → 98% accuracy