Design Specification

TASK

We have to build a 3 models to predict the interest rate

CLIENT

Rick, a person not willing to take risk while investing

APPROACH

After data cleaning and pre-processing, we have the features that can help us train our model that will predict the best interest rate. The final dataset on which we will be training our model has 157931 records covering all the possible scenarios based on the historical data. This data will help our model to learn and train itself to be able to predict the interest rate in future.

We have taken the approach of building 3 models and choosing the best model based on error metrics calculation (MAPE – Mean Absolute Percentage Error).

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

MAPE =

A_t = Actual Value

F_t = Forecasted or Predicted Value

Dependent Variable: - Target Variable i.e. Interest rate

Independent Variable: - (revol_bal, total_pymnt, loan_amnt, sub_grade, annual_inc, acc_now_delinq, delinq_2yrs, pub_rec, open_acc, inq_last_6mths, revol_util, emp_length, assr_state, application_type_JOINT,purpose_credit_card,purpose_debt_consolidation, purpose_educational,purpose_home_improvement, purpose_house, purpose_major_purchase,purpose_medical, purpose_moving, purpose_other, purpose_renewable_energy, purpose_small_business, purpose_vacation, purpose_wedding, term_ 60 months, home_ownership_MORTGAGE, home_ownership_NONE, home_ownership_OTHER, home_ownership_OWN, home_ownership_RENT, loan_income_ratio)

TRAIN AND TEST DATA SPLIT

We have divided our dataset into train and test data in the ratio 70% and 30% and trained our model. We calculated MAPE on this data split. Then we have implemented K-Fold cross validation for 5 splits and calculated the error metrics.

MODELS:

Linear Regression

We have designed our model with all the default parameters

LinearRegression – default parameters - fit_intercept=True, normalize=False, copy_X=True, n_jobs=None

The MAPE calculated on Test data is ~4.69 which infers that if the predicted interest rate is 10% the actual can vary from 10% +/- 0.049

The MAPE calculated on Train data is ~4.61
We performed 5-Fold cross validation on the same dataset with the same default parameters.
We calculated MAPE for each fold and then took the average to find out the MAPE for the model.

So, in our split, the conventional train test split yielded better result.

Random Forest

We have designed our model with all the default parameters of Random Forest

bootstrap=True, criterion='mse', max_depth=None,max_features='auto', max_leaf_nodes=None, min_impurity_split=1e-07, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1, oob_score=True, random_state=42, verbose=0, warm_start=False

The error metrics: -

MAPE for Test: 2.638090033898054 MAPE for Train: 1.0130208485029117 which we can consider for designing our final model.

We also trained our model on 5-Fold cross validation also.

- 5-Fold cross validation yielded MAPE on Test data: 3.990617695167303
- 5-Fold cross validation yielded MAPE on Train data: 0.973973908537018

Neural Network

We have designed our model with MLPRegressor algorithm of sklearn with all the default parameters

hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10

Error metrics Calculation: -

MAPE for Test: 3.25 MAPE for Train: 3.09 which we can consider for designing our final model.

We also trained our model on 5-Fold cross validation also.

- 5Fold cross validation yielded MAPE on Test data for test: 4.079

- 5Fold cross validation yielded MAPE on Train data: 2.93

CONCLUSION

From the 3 models that we have built, we can conclude that Random Forest yielded the best result with lowest MAPE score.

Split	Linear Regression	Random Forest	Neural Network
70:30 Split	Test – 4.69 Train – 4.61	Test – 2.63 Train – 1.01	Test – 3.25 Train – 3.09
5-Fold Split	Test - 4.079 Train – 4.614	Test – 3.99 Train – 0.97	Test - 4.079 Train - 2.93