Summary

 # Models to analyse customer churn at bank , Created 3 ML Models and comparison done and recommended better model. .

Link
https://github.com/RakeshRanjan14/Capstone_Project.git

NON-TECHNICAL EXPLANATION OF PROJECT
Customer Chrun is a major challenge for almost every industry. Specially for Banks this has a negative impact as bank's customer are generally long term and it take time to build that relation ship . Analysis of customer churn and reducing it by adopting different strategies is one of top priority for the banks. To Analyse customer churn at bank have build 3 models 1. Based on Logistic regression 2. based on Neural Network 3. Based on Decision Tree.

DATA
Data used is from Kaggle ,https://www.kaggle.com/datasets/bhuviranga/customer-churn-data. data has 10000 record , its an open data source .this data has features like credit score, balance , country , estimated salary , vintage of customer , credit card relationship and outcome customer churned or not.

MODEL

i created 3 models to do comparison and recommending a final model , these 3 models are based on 1. Logistic regression 2 Neural Network based on Tensorflow with different batch size epoch varying from 10 to 50 3. DecisionTreeClassifier with different max depth. Both Neural Network and high depth DT performed almost same but Tree depth with 5 may be overfitting scenario so recommended neural network with batch size 10 epoch 50.

HYPERPARAMETER OPTIMSATION

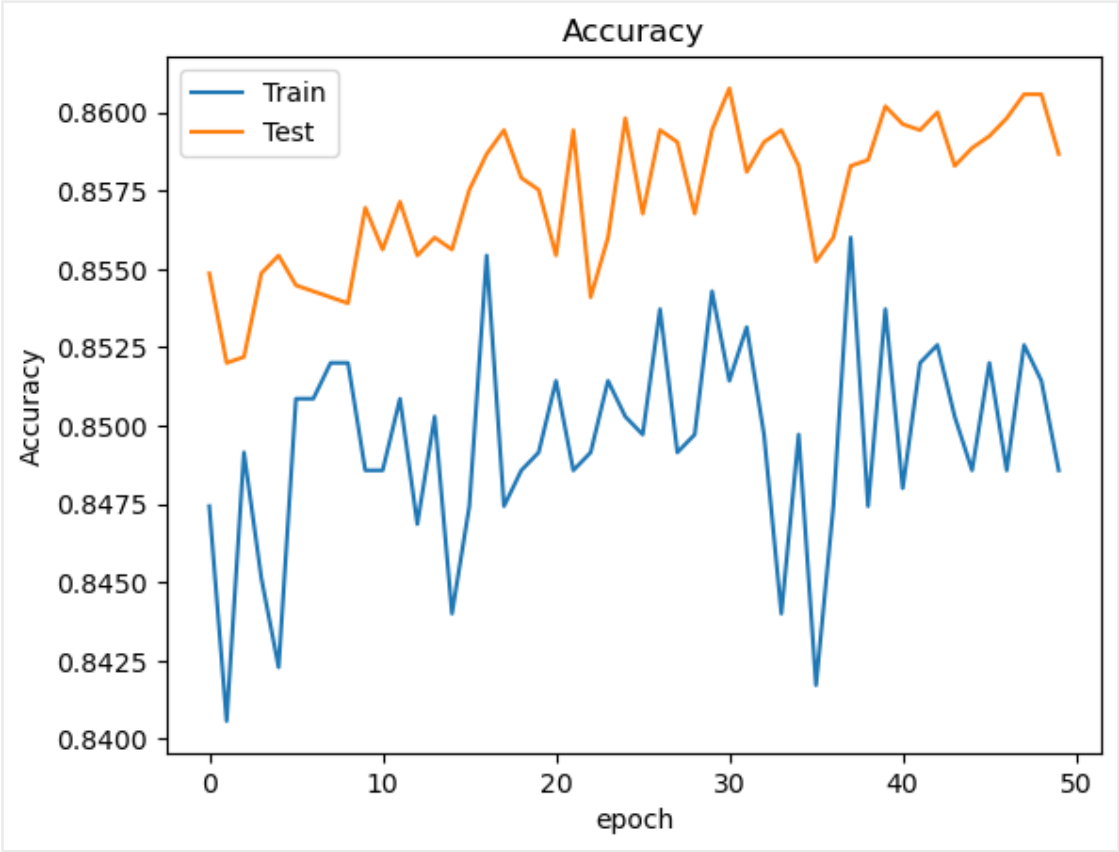
For Neural network Hyperparameter tuned were 1. Number of hidden layers, started with 11-8-4-1 (1 input layer , 2 hidden layers and 1 output) but settled with 1 input-1 hidden-and 1 output. 2. Number of nodes in each layer started with 11-6-4-1 but turn out to be 11-8-1 performed better 3. Epoch and Batch Size , epoch tuned from 10 to 50 and batch size tuned from 20 to 10.

RESULTS

Model 1 which is based on Logistic regression scored- Score of logistic

regression is 80.66%
Model 2 - Neural Network based model with different epoch ,batch size , different hidden layer and nodes scored between 83-86%
Model 3 - Decision Tree based model with different depth of default to 2 to 5 scored between 81-85.5%, bigger depth like 5 may lead to overfitting so recommended Model2.

Model 2 performance with different epoch



CONTACT DETAILS
https://github.com/RakeshRanjan14/Capstone_Project.git
Rakesh Ranjan -rakesh.ranjan@yahoo.com

Data

 # Datasheet Template
Data used is from Kaggle ,https://www.kaggle.com/datasets/bhuviranga/customer-churn-data. data has 10000 record , its an open data source .this data has features like credit score, balance , country , estimated salary , vintage of customer , credit card relatioship and outcome customer churned or not.

Motivation

Data was created for the purpose to analyse the churn pattern and develop strategies and interventions to reduce churn and improve customer retention

Author not known , so we can say ABC Bank.

The customer churn dataset is a collection of customer data that focuses on predicting customer churn, which refers to the tendency of customers to stop using a company's products or services. The dataset contains various features that describe each customer, such as their credit score, country, gender, age, tenure, balance, number of products, credit card status, active membership, estimated salary, and churn status. The churn status indicates whether a customer has churned or not. The dataset is used to analyze and understand factors that contribute to customer churn and to build predictive models to identify customers at risk of churning. The goal is to develop strategies and interventions to reduce churn and improve customer retention

Composition

its an open data source .this data has features like credit score, balance , country , estimated salary , vintage of customer , credit card relatioship and outcome customer churned or not.

Few of the data composition listed below

Sex
Male - 5457,
Female - 4543

Country
France - 5014,
Germany - 2509,
Spain - 2477

estSalary
max -199992.48,
min -11.58

Credit_Score
max -850,
min - 350

Output(Y) customer churned or not(Exited)

0 --7963- No,

1 --2037 -Yes

No Missing record for each column
ID_cu 0
Surname 0
Credit_score 0
Country 0
sex 0
Age 0
Tenure_with_bank_year 0
Balance 0
NumOfProducts 0
OwnCard 0
IsActiveMember 0
estSalary 0
Exited 0

Customer Surname and credit score is present in the data set but these 2 only will not contribute to confidentilty,

Collection process

Collection process is not known , but it seems it is been generated by bank and given as part of this analyis , this is taken from broader bank data set. colection period is also not known.

Preprocessing/cleaning/labelling

- Data had no missing record as such,
- To make it suitable for the all 3 models,string for Country and Sex need to be converted to numeric, created dummies for country and sex,
- country_num=pd.get_dummies(X["Country"],drop_first=True) and similarly created dummies for sex
Sex_num=pd.get_dummies(X['sex'],drop_first=True).
- As the values of Balance and est Salary was in large range so scalling done for overall data set using standard scaler -scaled_value = StandardScaler()
X_train = scaled_value.fit_transform(X_train)
X_test = scaled_value.transform(X_test)

Uses
Data set was created for churn analysis so should be used for this purpose only.
is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups- Not as

such

Distribution

Data fairly was normal distribution for almost all fields.
- Its open data source as mentioned in Kaggle.

Maintenance

- Not known.