Final Project Report
CHARITY DONORS PREDICTION

CHIRUMAMILLA SAIDARAO

04 - 27 - 2020

Business Understanding

Business Problem

There is one charity called charity which feeds homeless people, for donations this charity sends postal mails to residents of Delhi requesting to donate for a cause, from the historical data it is evident that residents who earn more than 50 thousand dollars per annum are more likely to donate. But the charity cannot be able to figure out how to send postal mails to those who most likely to donate to charity and avoid sending postal mails to those who are most likely are not going to donate to charity so that charity can save much money

Dataset

the dataset consists of 14 columns that describe the resident of Delhi, and the dataset consists of 45222 instances. I will split the data into training, and testing data use them to build a robust model for prediction. The dataset I found is from the below link

https://github.com/udacity/machine-learning/blob/master/projects/finding_donors/census.csv

Proposed Analytics Solution

The solution for the above problem is predicting whether the resident of Delhi is earning more or less than 50 thousand dollars per annum. So, I will use a few supervised models for prediction compare them in terms of metrics, and I will come up with the best model that suits for this problem PHASE-1

Data Exploration and Preprocessing

The dataset I used contains a total of 14 features and information about features are as follows

age: in years(continuous)

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked(Categorical)

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.(categorical)

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse(categorical)

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces(categorical)

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried(categorical)

race: Black, White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other(categorical)

sex: Female, Male.(categorical)

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

The above-described dataset contains categorical and continuous features with 45222 instances without missing values.

Descriptive features: - 'age', 'workclass', 'education_level', 'education-num','marital-status', 'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',

Target feature: - income

Data information

RangeIndex: 45222 entries, 0 to 45221
Data columns (total 14 columns):

#	Column	Non-NullCount	Dtype
0	age	45222 non-null	int64
1	workclass	45222 non-null	object
2	education_level	45222 non-null	object
3	education-num	45222 non-null	float64
4	marital-status	45222 non-null	object
5	occupation	45222 non-null	object
6	relationship	45222 non-null	object
7	race	45222 non-null	object
8	sex	45222 non-null	object
9	capital-gain	45222 non-null	float64
10	capital-loss	45222 non-null	float64
11	hours-per-week	45222 non-null	float64
12	native-country	45222 non-null	object
13	income	45222 non-null	object

dtypes: float64(4), int64(1), object(9)

memory usage: 4.8+ MB

Data Quality Report

The dataset contains continuous and categorical features I have divided features into two sets as continuous and categorical based on the datatype of the feature

Data Quality report table for Continuous features

Data Quality Report for Continuous features

	feature	count	missing values(%)	Unique values	minimum value	1ST Quartile	mean	median	3RD Quartile	maximum value	standard deviation
0	capital-gain	45222	0.0	121	0.0	0.0	1101.430344	0.0	0.0	99999.0	7506.430084
1	hours-per-week	45222	0.0	96	1.0	40.0	40.938017	40.0	45.0	99.0	12.007508
2	capital-loss	45222	0.0	97	0.0	0.0	88.595418	0.0	0.0	4356.0	404.956092
3	age	45222	0.0	74	17.0	28.0	38.547941	37.0	47.0	90.0	13.217870

Data Quality Report for Categorical Features

Data Quality Report for Categorical Features

	feature	count	missing values(%)	mode	mode frequency	mode percentage	2nd mode	2nd mode frequency	2nd mode percentage
0	workclass	45222	0.0	Private	33307	73.652205	Self-emp-not-inc	3796	8.394144
1	marital-status	45222	0.0	Married-civ-spouse	21055	46.559197	Never-married	14598	32.280748
2	sex	45222	0.0	Male	30527	67.504754	Female	14695	32.495246
3	education_level	45222	0.0	HS-grad	14783	32.689841	Some-college	9899	21.889788
4	race	45222	0.0	White	38903	86.026713	Black	4228	9.349432
5	relationship	45222	0.0	Husband	18666	41.276370	Not-in-family	11702	25.876786
6	occupation	45222	0.0	Craft-repair	6020	13.312105	Prof-specialty	6008	13.285569
7	income	45222	0.0	<=50K	34014	75.215603	>50K	11208	24.784397
8	native-country	45222	0.0	United-States	41292	91.309540	Mexico	903	1.996816

Missing Values and Outliers

Handling Missing values:-

Missing values can be handled using the deletion method or imputation method. There is a chance of losing data by using the deletion method so that imputation would be a good method for this dataset. So I have replaced missing variable by mean or mode for continuous or categorical features respectively

Handling Outliers Values:-

The values lower than $\mu - 2 * \sigma$ and higher than $\mu + 2 * \sigma$ are

considered as outliers and replaced with particular column mean

Normalization

Normalization is a technique usually applied in the process of data preparation in machine learning. The aim of normalization is to switch the values of numeric columns in the dataset to a standard scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have varying ranges.

I used MinMaxScaler Preprocessing technique using scikit learn library

```
X_scaled = scale * X + min - X.min(axis=0) * scale
where scale = (max - min) / (X.max(axis=0) - X.min(axis=0))
```

Feature Selection and Transformations

Feature Selection: -

10 best features are selected by using impurity-based univariate feature selection from 13 descriptive features. For the continuous feature, I used the threshold method to partition data and converts into categorical features

Before feature selection:-

age, workclass, education_level, education-num,
marital-status, occupation, relationship, race, sex,
capital-gain, capital-loss, hours-per-week, native-country

After Feature Selection:-

Capital loss, race, native country

Feature Transformation: -

I have used get_dummies for feature transformation from pandas library which transformed feature using one hot-coding

Before Feature Transformation there are 10 descriptive features, and these are transformed into 56 features using one-hot coding

Model Selection and Evaluation

Evaluation Metrics

The business problem is predicting whether the resident is earing more than 50 thousand dollars per annum, this prediction helps to decreases the cost of the charity to send postcards maximum donations. The accurate metric for this problem is Accuracy. Accuracy gives the prediction rate of the problem which is required for this problem high accuracy indicates the high prediction of residents who are

most likely earn more than 50 thousand dollars who are more likely to donate to charity

Models

I trained four models using preprocessed data, and the models are as follows

DecisionTreeClassifier(Information based learning)

KNeighborsClassifier(Similarity-Based Learning)

Naive Bayes(Probability-Based Learning)

Logistic regression(Error Based Learning)

DecisionTreeClassifier :-

The inputs for model are13 descriptive features and one target feature for training . after training model is provided with sampled test data without target feature to evaluate model performance

I have trained DecisionTreeClassifier with different Tree Depths and also with different impurity measures and noted and compared accuracies for Tree Depths and different impurity measures

KNeighborsClassifier :-

The inputs for model are13 descriptive features and one target feature for training . after training model is provided with sampled test data without target feature to evaluate model performance

The model was evaluated with different k values and noted their accuracies

Naive Bayes:-

The inputs for model are13 descriptive features and one target feature for training . after training model is provided with sampled test data without target feature to evaluate model performance and calculated accuracy.

Logistic regression:-

The inputs for model are13 descriptive features and one target feature for training . after training model is provided with sampled test data without target feature to evaluate model performance and calculated accuracy.

Sampling and Evaluation Settings

The total dataset was split into 40% testing, and 60% training data, 60% training data is used to train the model, and 40% testing data is used to test model performance and select the best model among four models described earlier

Hyper-parameter Optimization

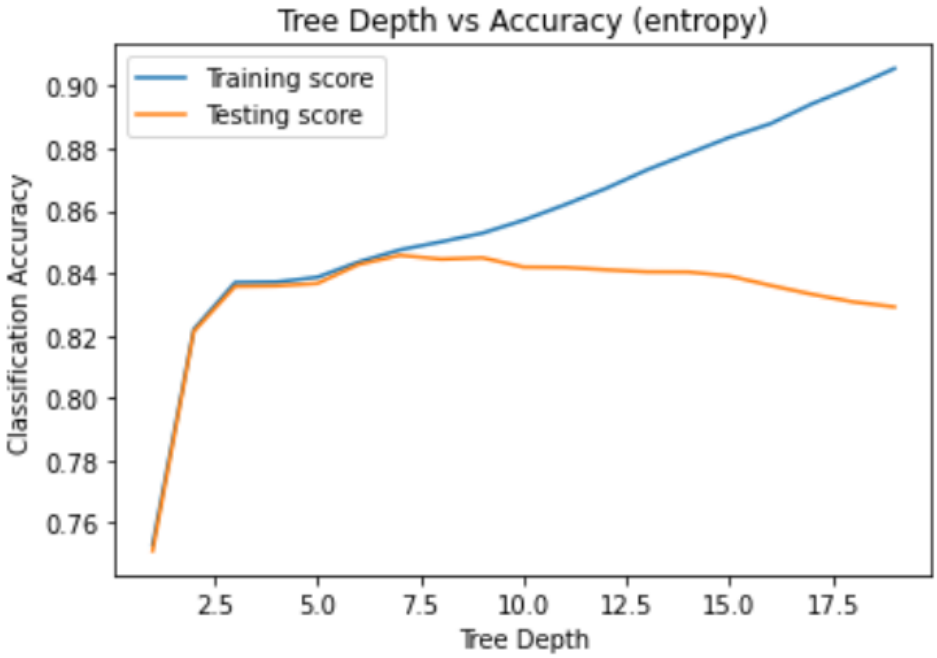
By tuning parameters of a model, we can increase the efficiency or performance of the model

As shown below, I have optimized parameters of decision tree and knn classifier to find the best performance with specific parameters

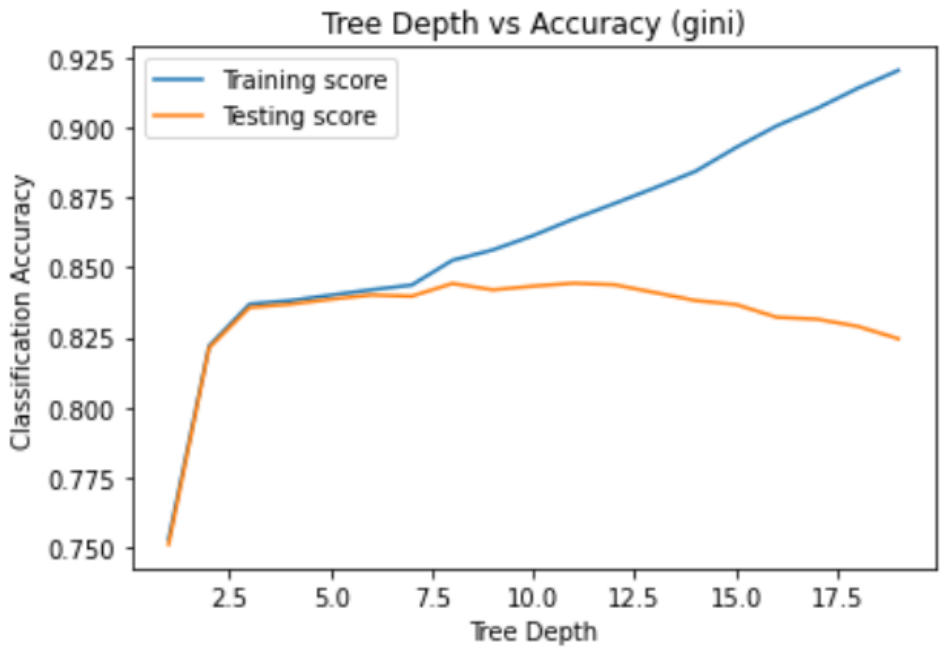
Evaluation

Performance of decision tree with different parameters as follows

```
[Text(0, 0.5, 'Classification Accuracy'), Text(0.5, 0, 'Tree Depth')]
```




```
[Text(0, 0.5, 'Classification Accuracy'), Text(0.5, 0, 'Tree Depth')]
```

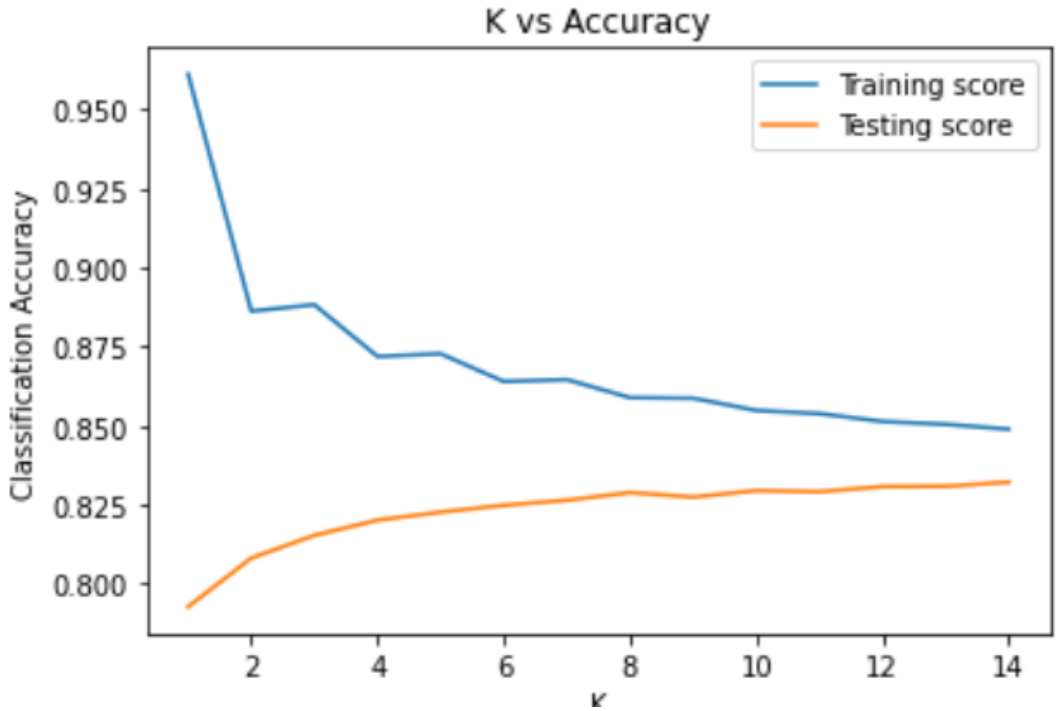


The final model relating to decision tree is the model with parameters tree depth = 7 impurity measure is Entropy because the has the highest testing score and good training score
.

The performance of KNN with different K values is as follows.

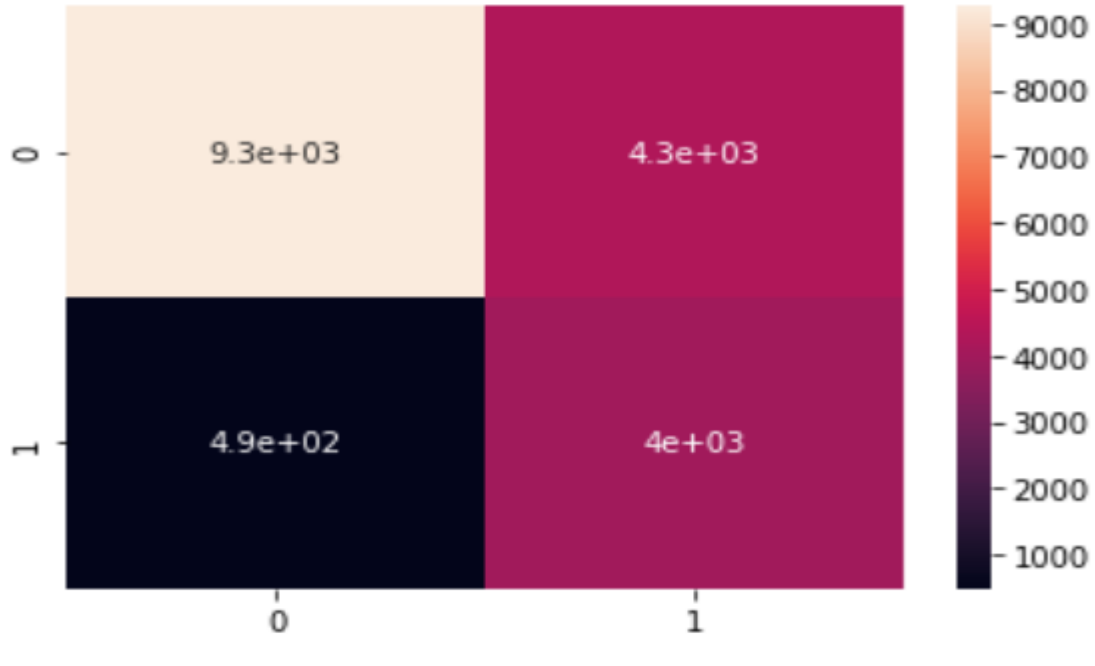
	K	Training score	Testing score
0	1	0.961265	0.792636
1	2	0.886227	0.808060
2	3	0.888254	0.815247
3	4	0.871890	0.820056
4	5	0.872738	0.822599
5	6	0.863966	0.824755
6	7	0.864556	0.826359
7	8	0.858843	0.828791
8	9	0.858659	0.827354
9	10	0.854789	0.829399
10	11	0.853794	0.829067
11	12	0.851288	0.830671
12	13	0.850404	0.830781
13	14	0.848856	0.832053

```
[Text(0, 0.5, 'Classification Accuracy'), Text(0.5, 0, 'K')]
```

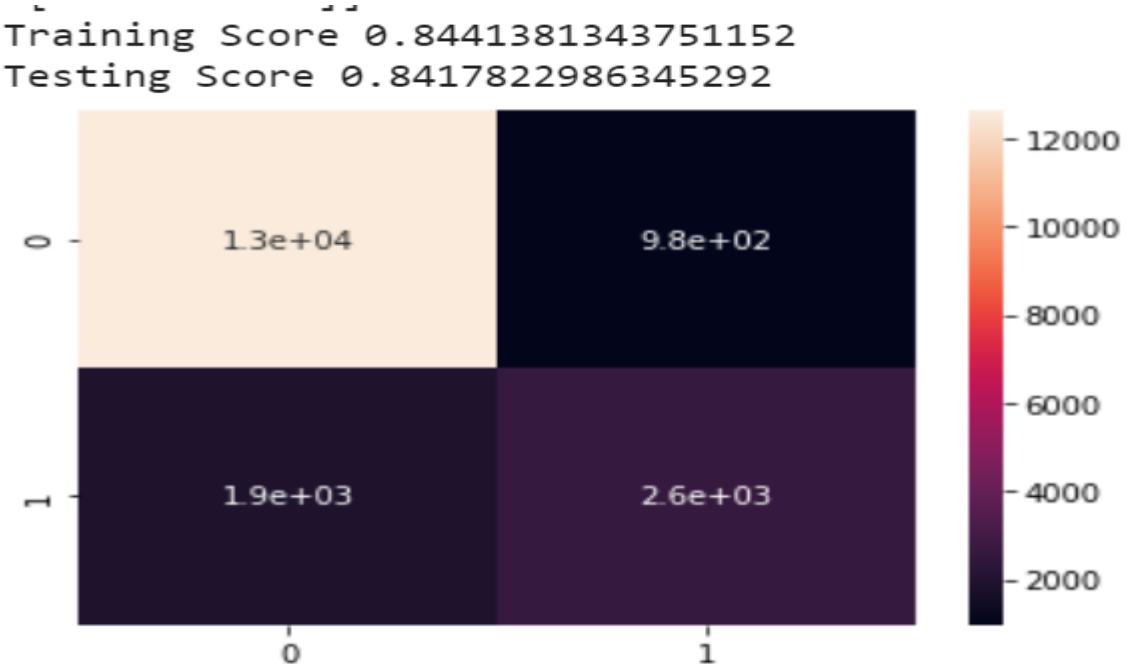


Performance of the Naive Bayes model as follows

```
[[9260 4325]
 [ 489 4015]]
Training Score 0.7345667637194561
Testing Score 0.7338714135662557
```



Performance of Logistic regression model as follows



Results and Conclusion

Results: -

MODEL	TRANING ACCURACY(%)	TESTING ACCURACY(%)
DECISION TREE	84.7	84.5
KNN	85.8	82.8
NAVIE BAYES	73.4	73.3
LOGISTIC REGRESSION	84.4	84.1

Conclusion:-

The best suitable model for the prediction problem is the decision tree model with tree depth = 7, and the impurity measure is Entropy.