ML 512: Course Project (Choice 2: explore datasets)

Venkata Subba Narasa Bharath, Meadam
venkatasubban.meadam@stonybrook.edu
SBU ID: 112672986

Algorithm: Decision Tree Algorithm

Datasets:

1. Credit Card Fraud Detection (Source: Kaggle)
2. The 20 newsgroups text dataset (Source: scikit-learn)
3. Adult Data Set (Source: UCI)

Deep Dive About Each Dataset:

Dataset 1: Credit Card Fraud Detection

Task

It is important that credit card companies can recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.
The datasets contain transactions made by credit cards in September 2013 by European cardholders.

Identify fraudulent credit card transactions.

Samples: 284807
Features: 31

Uniqueness
Highly Unbalanced Dataset

	Positives (Fraudulent Transaction)	Negatives (Non-Fraudulent Transactions)
Dataset	0.17%	99.83%

✓

Result:

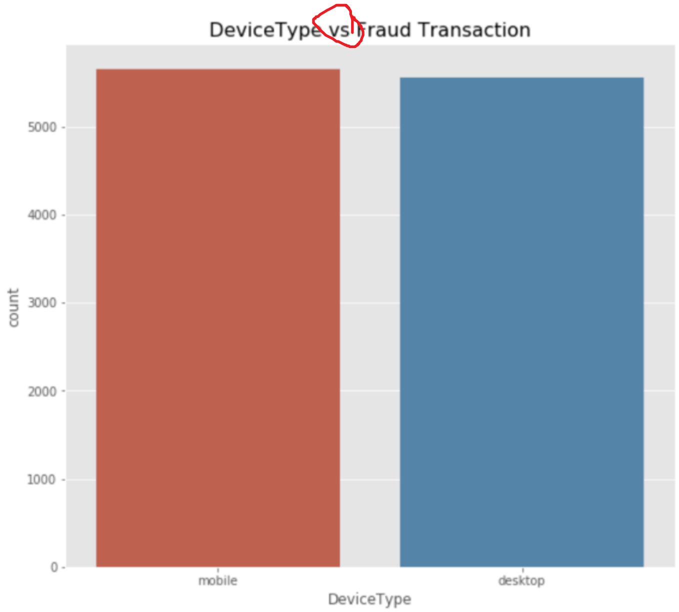
Evaluation Metric	Score
ROC AUC	0.88
Accuracy Score	0.99
F1-Score	0.8

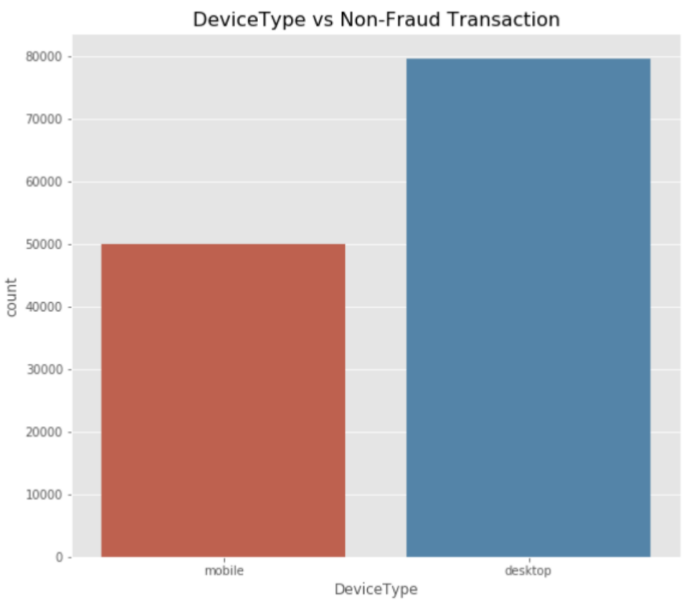
This is test score, right?

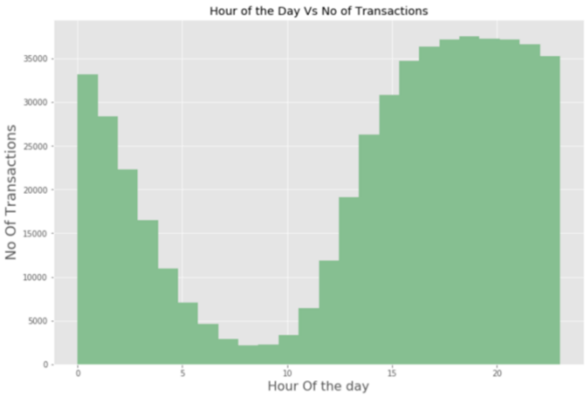
Observations/Comments:

1. The Data is Right-skewed as most of the transactions are of small amounts
2. Even though the number of transactions on mobile and desktop is almost the same, there are more fraudulent transactions recorded on mobile compared to desktop. We can guess maybe more fraudulent transactions occur on mobile.

? Is that the right word here?







3. There is a clear pattern in several transactions that occurred during different hours of the day and as one would expect the transactions were few in the early hours of the day and peaked during the evening as one would expect.

4. In this task, we have seen how a decision-tree classifier performed on a highly imbalanced dataset (Imbalance ratio 99.83:0.17)

Would you consider this good or bad?

Dataset 2: The 20 newsgroups text dataset

Task

The classification of 20_newsgroup dataset is a supervised classification problem, there is news of 20 categories, each piece of news belongs to one category, the goal is to extract proper features and build an effective model to assign each piece of news to the correct category. Identify fraudulent credit card transactions. ?

What are the features?

Samples: 2034
Features: 190

Uniqueness
Sparsity

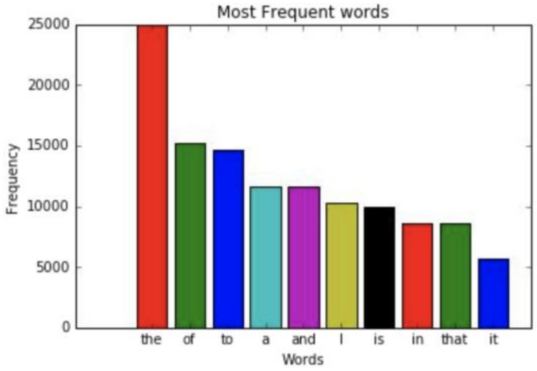
Sparsity	99.53%

Result:

Evaluation Metric	0.56
precision recall	0.54
F1-Score	0.55

Observations/Comments:

1) The total number of unique words can be 671564, a very huge number if we treat all of them as features, we need to extract only a proper proportion



How does this impact the models?

2) In this task, we have seen how a decision-tree classifier performed on a highly sparsed dataset (Imbalance ratio 99.53% sparsity)

Dataset 3: Adult Data Set

Task

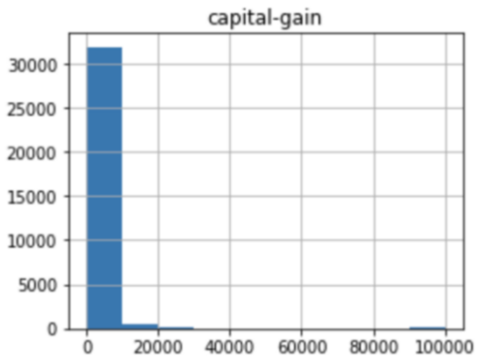
This data was extracted from the 1994 Census bureau database. *The prediction task is to determine whether a person makes over $50K a year*.

Samples: 32561
Features: 15

Result:

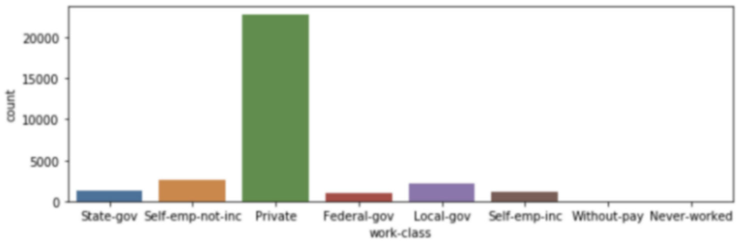
Evaluation Metric	Score
Accuracy Score	0.80
F1-Score	0.62

Observations/Comments:



1. People do not usually have savings other than their daily income. There are very few individuals who invest, however, and there are just a tiny number of outliers who earn more than 90000 from capital gains. Nonetheless, the average loss tends to be about 2000 for the people who had a capital loss

2. Most citizens work in the private sector, and the remainder is divided equally between state-gov, federal-gov, local-gov self-emp-inc, and self-emp-not-inc.



ok

Being a Consultant - Why do I care?

1. **Great Intuition**
   Decision trees give us an easy-to-understand algorithm which is easy to interpret for a wide range of audience, this has been seen for all the three datasets. (we are not talking about accuracy; we are only concerned with how decisions are made at each split)
2. **Followed a good metric**
   We followed the standard metrics of F1-Score/Accuracy/Precision to evaluate our models. This gives value to each of our models.
3. **Decision Tree is not ~One Size fits all**
   From, the results we can see that decision tree is not the de-facto choice for all the types of datasets, even though all our 3 datasets are about classification. The results were different in each of them, we need to see how our data is distributed, before fitting into a Decision tree Classifier.

Being a Critic – How can I Improve?

1. **Need datasets with both Categorical and Continuous Variables:**
   All the three datasets were either completely categorical (or) numerical, even though the end goal was classification, it would have been better if there were few datasets which had both type of features, that would help us know more about the decision tree.
2. **Overfitting/Unstable/Higher training time**
   One common feature, we observed for all the three datasets was that Decision trees highly overfit and are volatile to small changes in. the dataset. They are computationally expensive to train as compared to other algorithms.

References:

1. https://www.kaggle.com/uciml/adult-census-income
2. https://medium.com/district-data-labs/building-a-classifier-from-census-data-18f996c4d7cf
3. https://www.kaggle.com/mlg-ulb/creditcardfraud
4. https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets
5. https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html
6. https://medium.com/themlblog/text-classification-using-machine-learning-cff96602c264
7. https://archive.ics.uci.edu/ml/datasets/adult

This one is obviously much weaker than the previous one, but it is interesting and hits all the minimum criteria. Good job with first pass analysis, but obviously there's much more to do. (Why does different data perform differently on decision tree? Are the splits different? Is overfitting behavior different? How good is "good enough" in each application, and is DT the best tool to get there?)

min criteria: 5
writing: 0.5
interestingness: 0.5