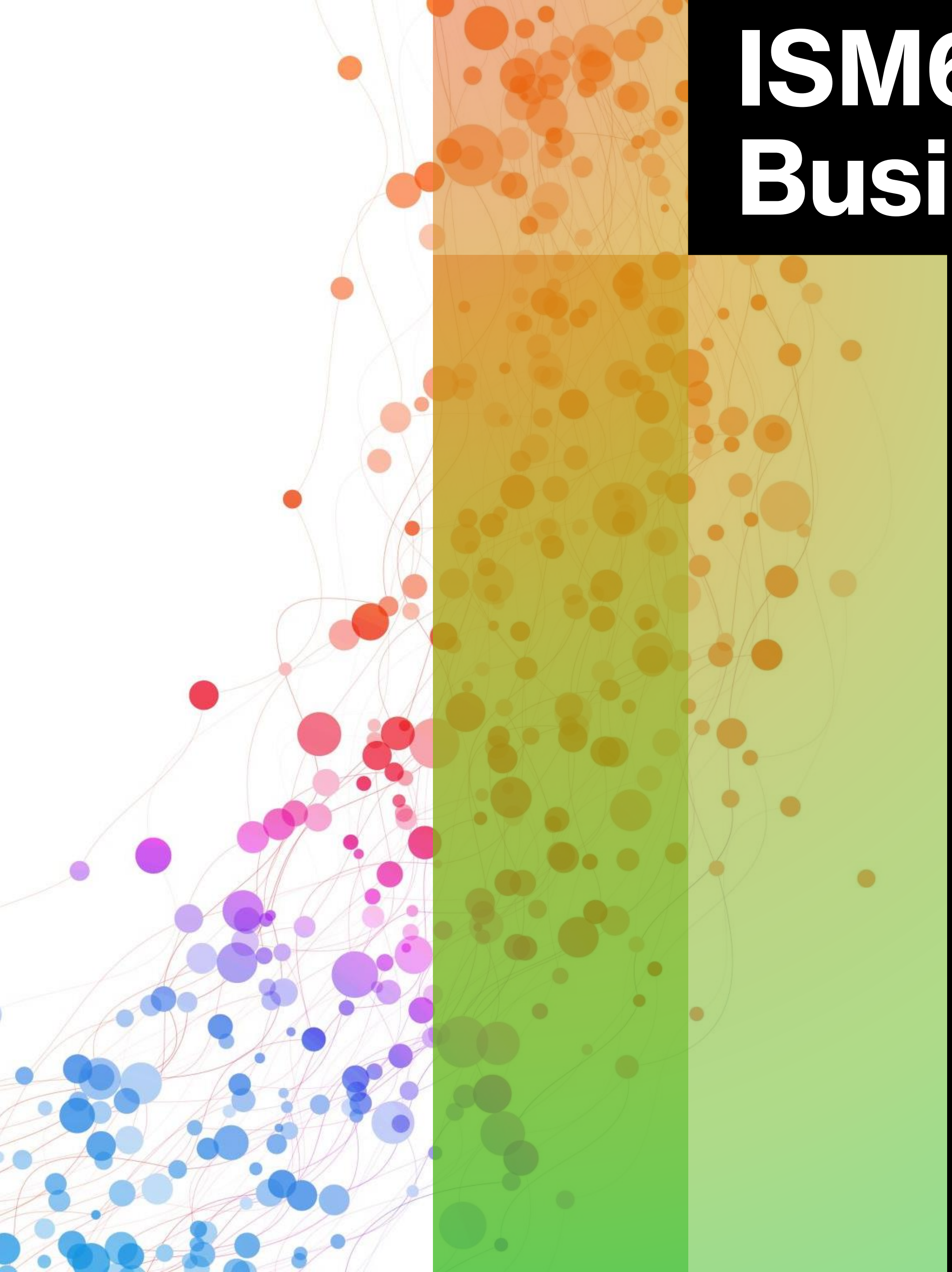


ISM6562 Big Data for Business Final Project

Identifying prospective customers from Census Data

TEAM ZENITH

Ganti Uday

Rahul Reddy Vemparala

Sri Kumar Dundigalla

Nikhil Vuppalavanchu

Description of the business problem

- Motive

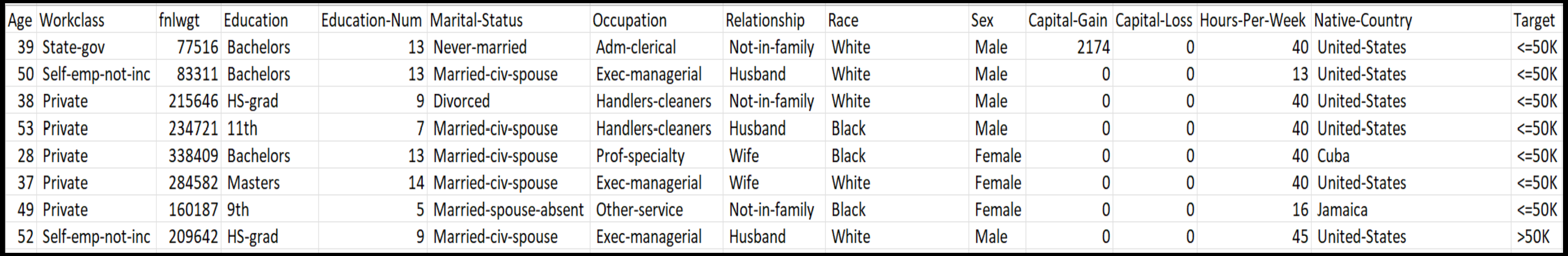
[To identify if a person could be a potential customer]

- How may a data analysis/modeling project address this challenge/problem?

[We can utilize publicly available data such as Census and ML Algorithms to best identify potential customers.]

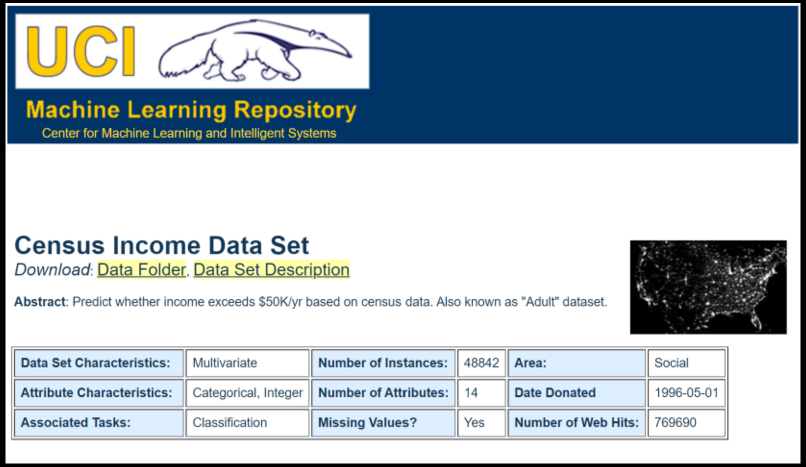
Data Description

- Description of the dataset used



Age	Workclass	fnlwgt	Education	Education-Num	Marital-Status	Occupation	Relationship	Race	Sex	Capital-Gain	Capital-Loss	Hours-Per-Week	Native-Country	Target
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K

- This data was extracted from the census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html
  - Extraction was done from the 1994 Census database



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

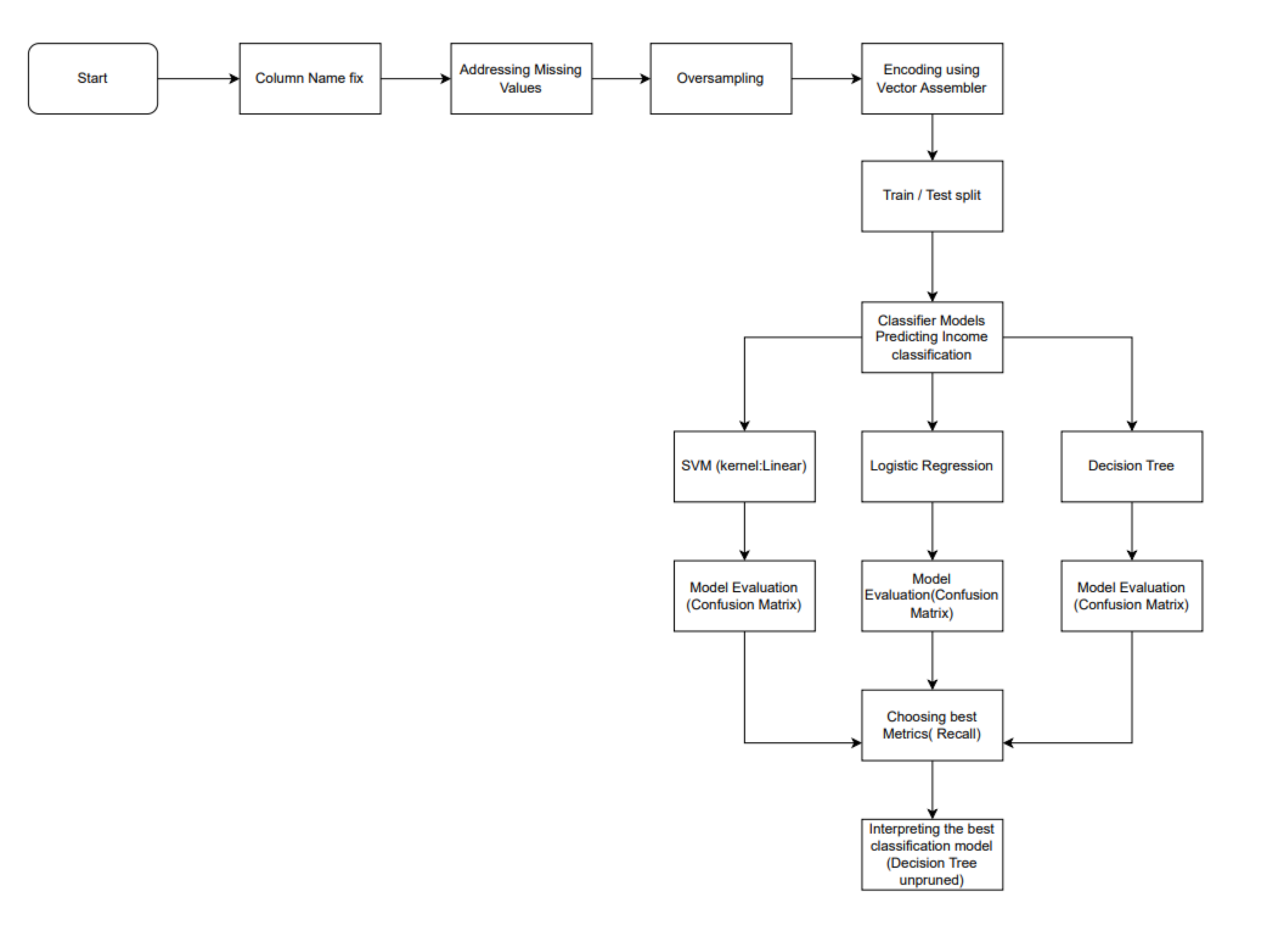
Census Income Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	769690

Pictorial representation of the Data Flow Model



Start

Column Name fix

Addressing Missing Values

Oversampling

Encoding using Vector Assembler

Train / Test split

Classifier Models Predicting Income classification

SVM (kernel:Linear)

Logistic Regression

Decision Tree

Model Evaluation (Confusion Matrix)

Model Evaluation(Confusion Matrix)

Model Evaluation (Confusion Matrix)

Choosing best Metrics(Recall)

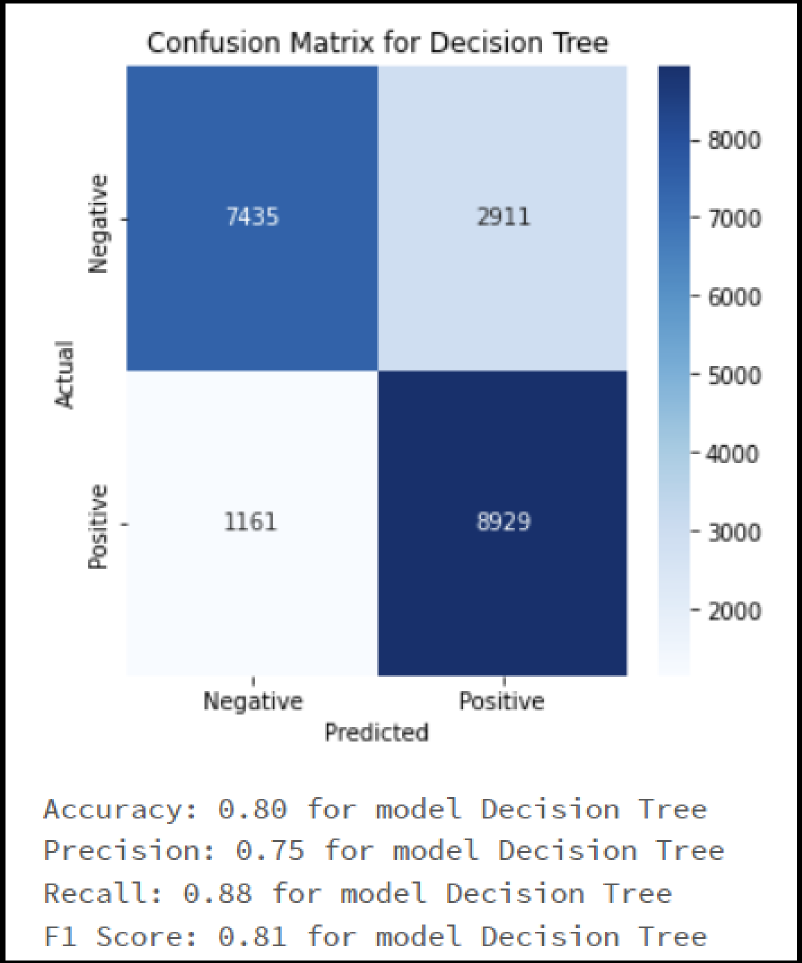
Interpreting the best classification model (Decision Tree unpruned)

Now, Let’s check out the code!

Cost Estimates for Errors

- Evaluating Various Errors:
- 1. True Positive
- 2. True Negative
- 3. False Positive
- 4. False Negative

Picking the Best Model



```
Accuracy: 0.80 for model Decision Tree
Precision: 0.75 for model Decision Tree
Recall: 0.88 for model Decision Tree
F1 Score: 0.81 for model Decision Tree
```

Addressing the Challenges faced

- Hyper parameter tuning, -parallelize using spark
- Data Imbalance
- Missing Data – categorical columns

Future Scope

- More features imply a more robust model(less variance).
- Performance tuning: Random, and Exhaustive search
- Scope for a regression problem.