UC Berkeley Data Analytics
GROUP 3

Divorce Predictor Using Data Modelling & Machine Learning

Priyanka Senapati
Rajath Narasimha
Thi Nguyen

March 26th 2020

OVERVIEW

Divorce Predictor Project

- Motivation for the Topic
- Challenges in Datasets & Correlations
- Description of the Data Source
- Description of the Analysis
- Model Training Evaluation Analysis
- Technologies, Languages, Tools, Algorithms
- Results of the Analysis
- Anything we wished to do differently?

Motivation for the Topic

- Curious to find out why over 50% of all marriages in the US end up in a divorce or separation.
- What circumstances lead to a divorce? How to take precautions in a holistic way?
- Does occupation, age, race & location play a role in the divorce?
- Using our Data Analytics & Machine Learning skills we were able to find answers.
- UC Irvine real survey Divorce Predictor data set with sample size 170 with 54 attributes.
- Our quiz which will predict your divorce rate. Caution, take it at your own risk!

Challenges in Datasets & Correlations

- Biggest challenge to find a good meaningful sample size with real life surveys.
- Different file formats, .csv, .pdf, .html to process.
- Training data set vs. The size of the out of sample test data set correlations.
- We assume people's behaviors would be similar across diverse backgrounds.
- Unable to extrapolate data sets.
- May result in high variance.

Description of the Data Source

Data Acquisition:

1. **Data for machine learning**

   UCI divorce predictor dataset (csv)

2. **U.S. Census Data:**

   United States marriage and divorce rate (pdf)

   Marriage rate by state (pdf)

   Divorce rate by state (pdf)

3. **Peer reviewed article:**

   The growing Racial and Ethnic Divide in U.S. Marriage Patterns – Future Child (pdf)

4. **Online research:**

   Divorce rates by occupation and salary, aggregates (google spreadsheets from flowingdata.com)

Table 1

Women's Age-Specific Rates of First Marriage and Divorce by Race, Ethnicity, and Nativity

	Panel A. Marriage						
Age	White	Black	Asian/Pacific Islander	American Indian/Native Alaskan	Hispanic, Total	Hispanic, U.S. born	Hispanic, foreign born
15–19	8.7	5.0	8.5	20.3	16.7	13.1	32.6
20–24	58.9	23.0	41.4	53.5	59.1	50.4	81.3
25–29	115.6	43.0	133.7	76.6	81.0	75.9	89.2
30–34	130.6	47.6	152.5	74.9	87.4	83.0	92.1
35–39	123.0	44.6	129.1	70.5	80.4	72.7	86.8
40–44	111.6	39.4	100.5	51.8	77.9	72.6	82.2
	Panel B. Divorce						
Age	White	Black	Asian/Pacific Islander	American Indian/Native Alaskan	Hispanic, Total	Hispanic, U.S. born	Hispanic, foreign born
20–24	48.44	40.13	12.23	63.61	26.79	36.74	16.13
25–29	38.80	44.29	13.23	52.02	26.71	40.43	15.31
30–34	31.60	44.43	15.95	40.15	25.03	37.09	16.83
35–39	29.66	41.20	12.98	41.58	23.70	36.31	16.43
40–44	26.33	38.86	13.07	48.60	21.47	30.15	16.78

```
# Pivot table
pivot_marriage_race = pd.pivot_table(melt_marriage_race, values='Marriage_Rate', index=['Age'],
                     columns=['Race'], aggfunc=np.sum)
pivot_marriage_race
```

Race	American Indian/Native Alaskan	Asian/Pacific Islander	Black	Hispanic, Total	Hispanic, U.S. born	Hispanic, foreign born	White
Age							
15–19	20.3	8.5	5.0	16.7	13.1	32.6	8.7
20–24	53.5	41.4	23.0	59.1	50.4	81.3	58.9
25–29	76.6	133.7	43.0	81.0	75.9	89.2	115.6
30–34	74.9	152.5	47.6	87.4	83.0	92.1	130.6
35–39	70.5	129.1	44.6	80.4	72.7	86.8	123.0
40–44	51.8	100.5	39.4	77.9	72.6	82.2	111.6

Description of the Data Source

Data Transformation:

The datasets were cleaned up via jupyter notebook and each saved into a csv file for later use. Steps

Step 1: Read file (pd.read_csv, pd.read_html, read_pdf using tabula-py)

Step 2: Drop null columns and Reset column names

Step 3: Clean up data

Step 4: Change data types if needed

Step 5: Change to pivot if needed

Step 6: Save to csv file or postgreSQL for later use

Marriage and Divorce Rate by Race

```
1 dfs = read_pdf('nihms777225.pdf', pages='all', multiple_tables=True)
2 dfs[0]
```

44]:

	0	1	2	3	4	5	6	
0	Age	White	Black	Asian/Pacific Islander	American Indian/Native Alaskan Hispanic, Total	NaN	Hispanic, U.S. born	Hi
1	15–19	8.7	5.0	8.5	20.3 16.7	NaN	13.1	
2	20–24	58.9	23.0	41.4	53.5 59.1	NaN	50.4	
3	25–29	115.6	43.0	133.7	76.6 81.0	NaN	75.9	
4	30–34	130.6	47.6	152.5	74.9 87.4	NaN	83.0	
5	35–39	123.0	44.6	129.1	70.5 80.4	NaN	72.7	
6	40–44	111.6	39.4	100.5	51.8 77.9	NaN	72.6	
7	NaN	NaN	NaN	NaN	Panel B. Divorce	NaN	NaN	
8	Age	White	Black	Asian/Pacific Islander	American Indian/Native Alaskan Hispanic, Total	NaN	Hispanic, U.S. born	Hi
9	20–24	48.44	40.13	12.23	63.61 26.79	NaN	36.74	
10	25–29	38.80	44.29	13.23	52.02 26.71	NaN	40.43	

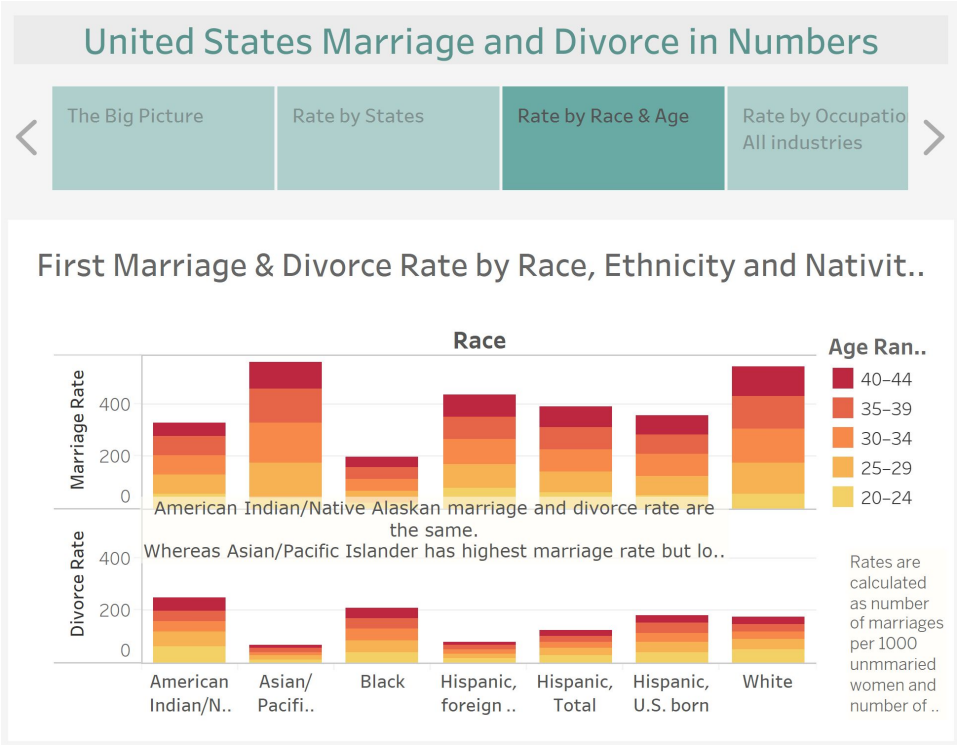
Description of the Analysis

Data Analysis:

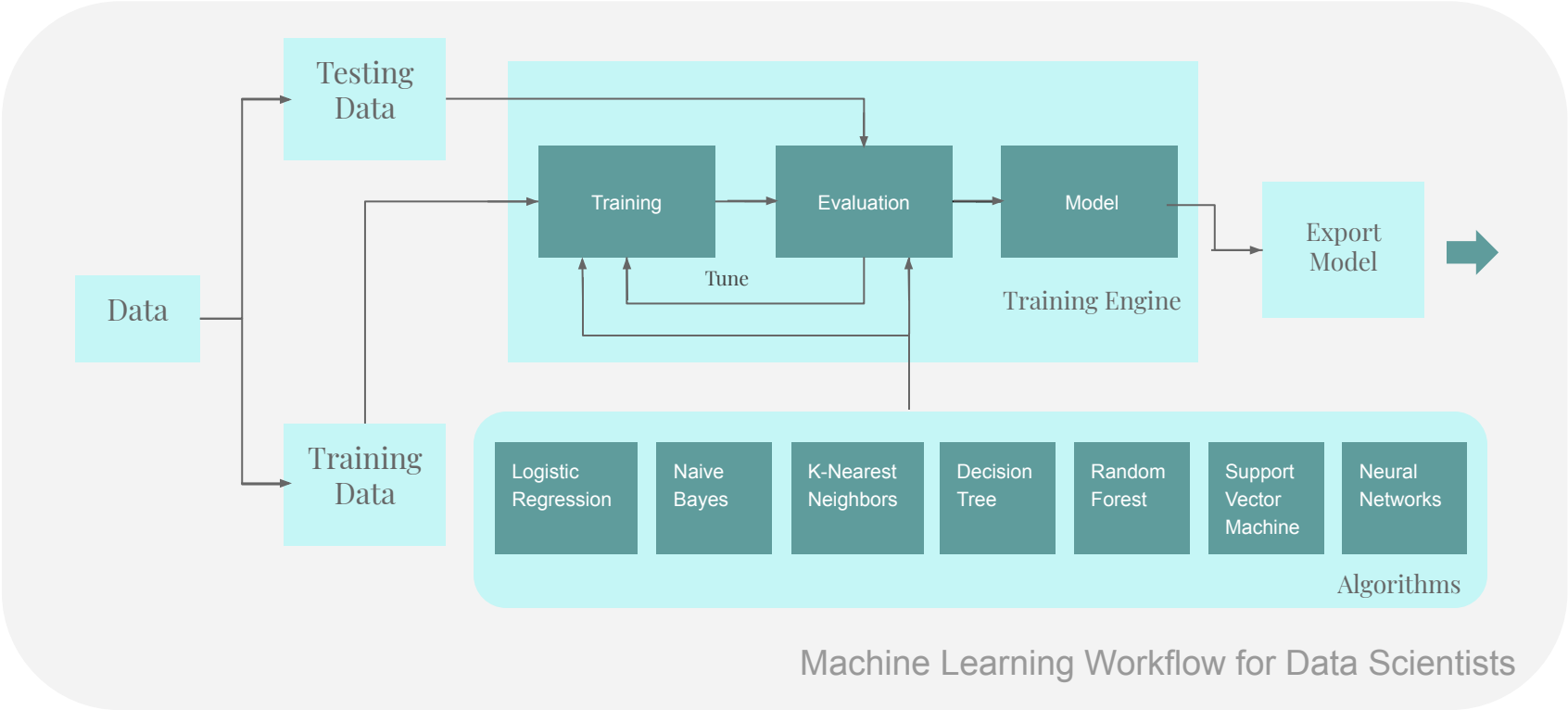
1. Perform analysis on tableau
2. The story is then published and embedded in home page of Flask app
3. Live demo of the app

Tableau Story:

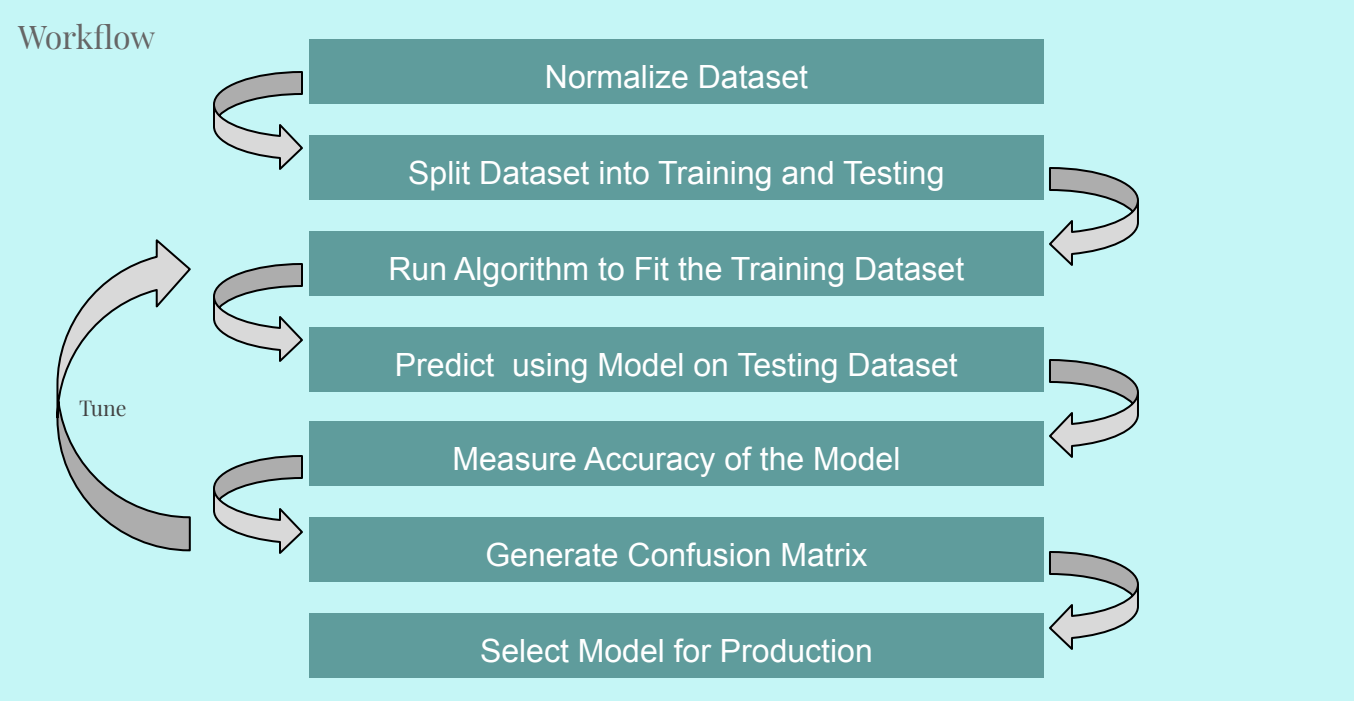
https://public.tableau.com/profile/thi7884#!/vizhome/nationalrates/Story_US_Marriage_Divorce



Model Training Evaluation Analysis



Model Training Evaluation Analysis



Model Training Evaluation Analysis

Machine Learning Algorithms

Logistic Regression Model selected for final Production Inferencing.

- *Logistic Regression Algorithm (***sklearn.linear_model.**LogisticRegression)

```
# Logistic Regression Algorithm
from sklearn.linear_model import LogisticRegression
lg = LogisticRegression(random_state=0,solver = "liblinear")
# Create Model
lg.fit(X_training,y_training)
# Predict outcome using the Attribute values from the testing dataset
y_predict = lg.predict(X_testing)
# Compute Accuracy of the model
print("Accuracy = ",((np.sum(y_predict==y_testing)/y_testing.shape[0])*100),"%",sep="")
```

```
Accuracy = 100.0%
```

```
print (y_testing)
```

```
[0 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1
 0 0 0 0 0 1]
```

```
print (y_predict)
```

```
[0 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1
 0 0 0 0 0 1]
```

Model Training Evaluation Analysis

Machine Learning Algorithms

- *Naive Bayes Algorithm (***sklearn.naive_bayes.**GaussianNB)
- *K-Nearest Neighbors (KNN) Algorithm: (***sklearn.neighbors.**KNeighborsClassifier*)*
- *Decision Tree Algorithm (***sklearn.tree.**DecisionTreeClassifier*)*
- *Random Forest Algorithm (***sklearn.ensemble.**RandomForestClassifier)*
- *Support Vector Machine (SVM) Algorithm (***sklearn.svm**.SVC*)*
- *Neural Networks Algorithm (***keras.models**.Sequential*)*

Sample Confusion Matrix

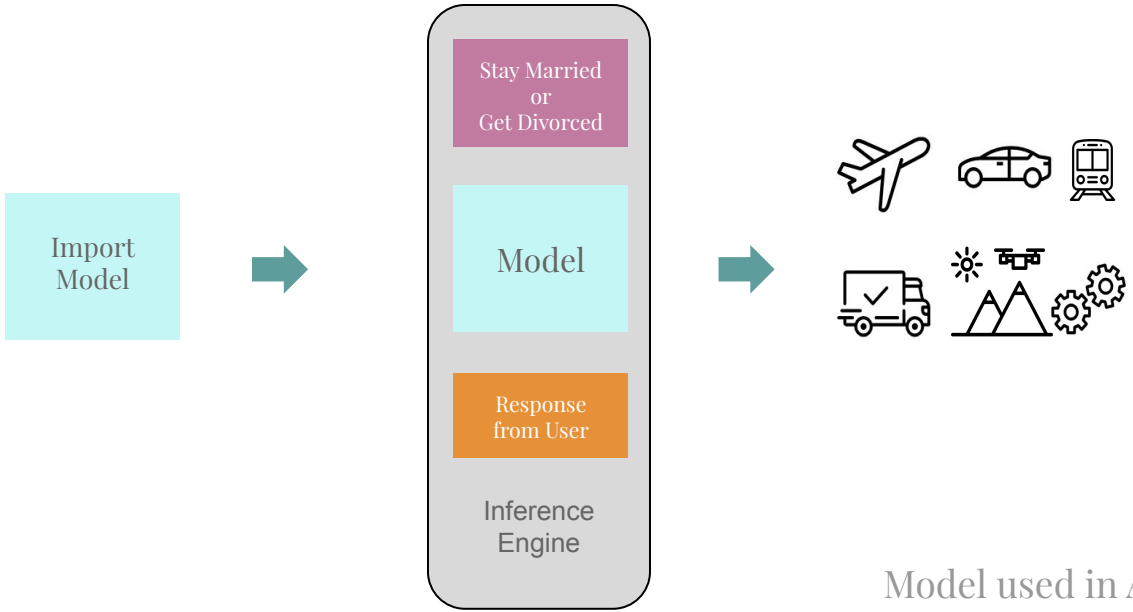
```
# Further enhancement: Rank the algorithms according to their accuracy metrics.
#
# Compute the Confusion Matrix based on the actual and predicted divorce outcomes
# Find the accuracy score taking the actual and predicted values
# Generate a Classification report
#
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
actual = y_testing
predicted = y_predict
results = confusion_matrix(actual, predicted)
print('Confusion Matrix :')
print(results)
print('Accuracy Score :',accuracy_score(actual, predicted))
print('Report : ')
print(classification_report(actual, predicted))
```

```
Confusion Matrix :
[[20  0]
 [ 1 22]]
Accuracy Score : 0.9767441860465116
Report :
              precision    recall  f1-score   support

           0       0.95      1.00      0.98        20
           1       1.00      0.96      0.98        23

    accuracy                           0.98        43
   macro avg       0.98      0.98      0.98        43
weighted avg       0.98      0.98      0.98        43
```

Inferencing in the Field



Model used in Applications

Results of the Analysis

- Actual Divorce cases have decreased significantly over past 2 decades.
- Low marriage rates might be a big contributing factor to it.
- Age: Lowest% 59+ , Highest% 40-49.
- Race: Lowest % Asian/Pacific Islanders VS Highest % American Indian, Alsakan
- State: Lowest % Iowa, Highest % Nevada.
- Occupation: Lowest % Chemical Engineers VS Highest % Dancers/Bartenders.

Anything we wished to do differently?

- More survey results with good sample size to predict correlations between:

a) GDP/Country/Religion/Spiritual Beliefs

b) Parental History & Dating History

c) Addictions, Fitness, Physical & Mental health

d) Income Levels, Prenups

e) Myers-Briggs Personality Data

Technologies, Languages, Tools & Algorithms

- Python Jupyter Notebook for data transformation ML
- HTML5 & CSS for local host website display & design
- Flask App in python, Visual Studio Code
- Tableau Public for visualization
- .csv data files for post processing

References

UCI divorce predictor dataset
https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set

U.S Census data
https://www.cdc.gov/nchs/nvss/marriage-divorce.htm

Marriage/Divorce rate by Ethnicity and Nativity
https://www.ncbi.nlm.nih.gov/pubmed/27134512

Divorce rate by Occupation
https://flowingdata.com/2017/07/25/divorce-and-occupation/

Online search
https://www.wf-lawyers.com/divorce-statistics-and-facts/

Thank You!