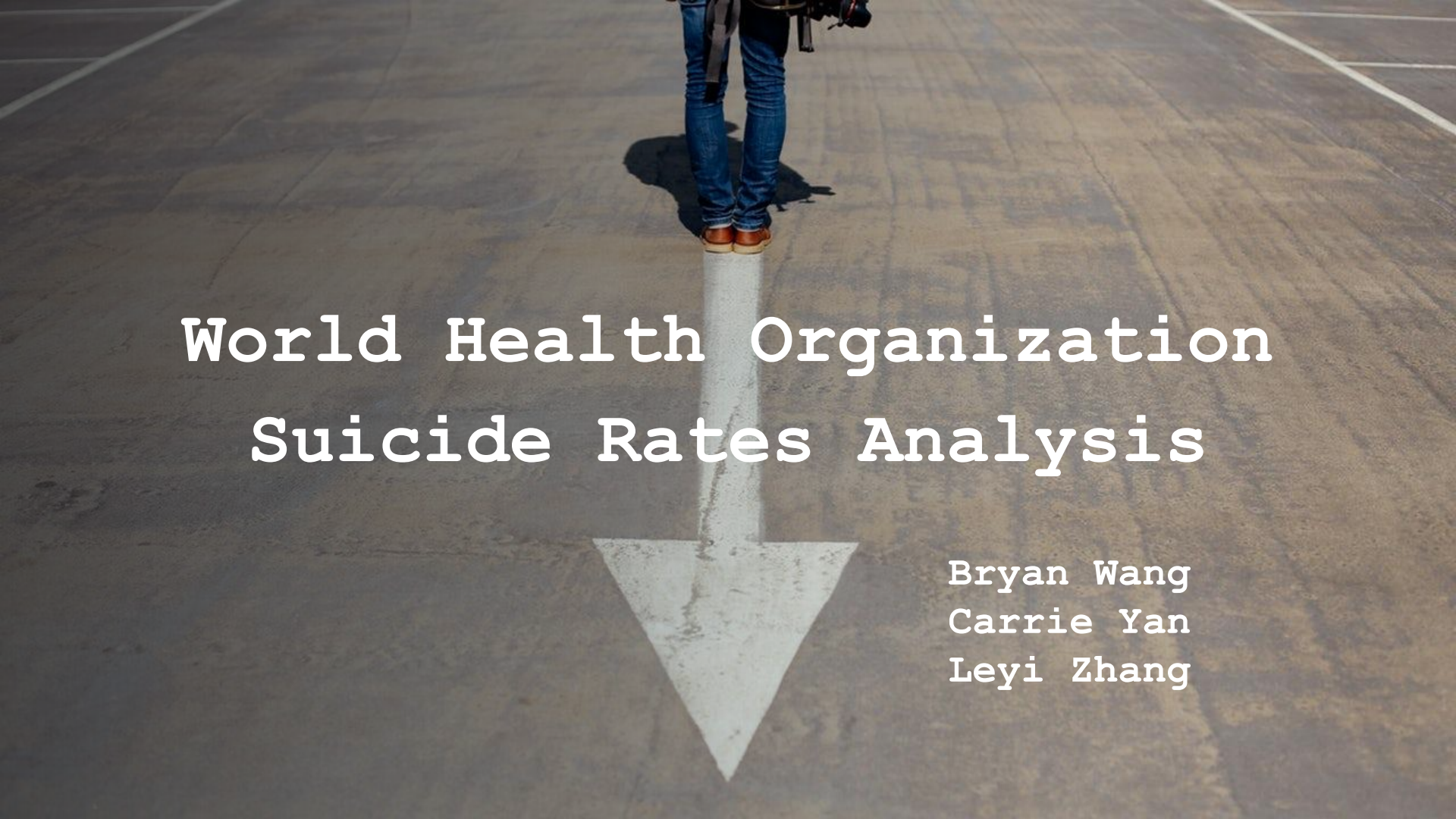


World Health Organization Suicide Rates Analysis

Bryan Wang
Carrie Yan
Leyi Zhang

Outline

1. Introduction
2. Research Question
3. Data Preprocessing & Preparation
4. Exploratory Data Analysis
5. Model Construction
6. Model Evaluation
7. Conclusion and Future Studies

NATIONAL SUICIDE PREVENTION LIFELINE
1-800-273-TALK (8255)
suicidepreventionlifeline.org

1.

Introduction

Let's talk about suicide

Key facts

- Close to 800 000 people die due to suicide every year.
- For every suicide there are many more people who attempt suicide every year. A prior suicide attempt is the single most important risk factor for suicide in the general population.
- Suicide is the second leading cause of death among 15–29-year-olds.
- 79% of global suicides occur in low- and middle-income countries.
- Ingestion of pesticide, hanging and firearms are among the most common methods of suicide globally.

Dataset

```
suicide.head()
```

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

- 12 variables, 27.8k observations
- Over the years from 1985-2016
- Kaggle dataset compiled from WHO and World Bank DataBank
- Find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum

2. Research Questions

- **Gain insights of global suicide rate over years through exploratory data analysis**
- **Build models to predict the future suicide rate for the six WHO world regions**
- **Choose the winner model to make the most precise prediction**

World Health Organization

3. Data Preprocessing and Preparation

```
df.printSchema()
```

```
root
 |-- country: string (nullable = true)
 |-- year: long (nullable = true)
 |-- sex: string (nullable = true)
 |-- age: string (nullable = true)
 |-- suicides_no: long (nullable = true)
 |-- population: long (nullable = true)
 |-- suicides/100k pop: double (nullable = true)
 |-- country-year: string (nullable = true)
 |-- HDI for year: double (nullable = true)
 |--  gdp_for_year ($) : string (nullable = true)
 |-- gdp_per_capita ($): long (nullable = true)
 |-- generation: string (nullable = true)
```

a. **Check Missing Values**
b. **Check data types**
   --- string type for numeric features
c. **Check duplicate features**
   --- provides no extra information

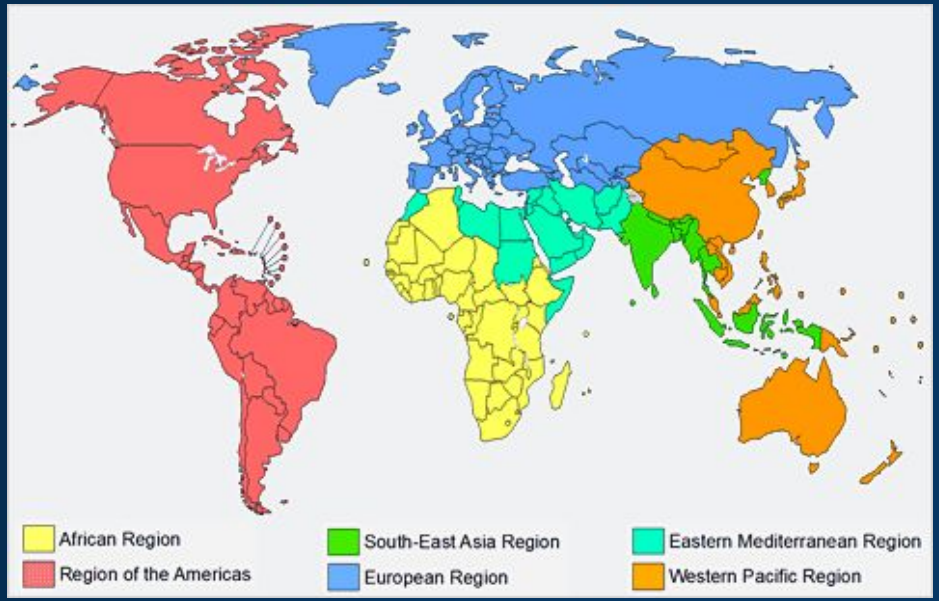
Lecture Notes 05: Spark SQL and DataFrames

3. Data Preprocessing and Preparation

a. Feature Engineering

141 Countries → #6 WHO World Regions

1985 -2016 → #10 3-year span



3. Data Preprocessing and Preparation

b. Dropped "country-year" since it provides no extra information

c. Deleted "HDI for year" since it contains about 70% missing values

d. Renamed features that contains " \", " ", "()" with underscores

Exploratory Data Analysis

3. Data Preprocessing and Preparation

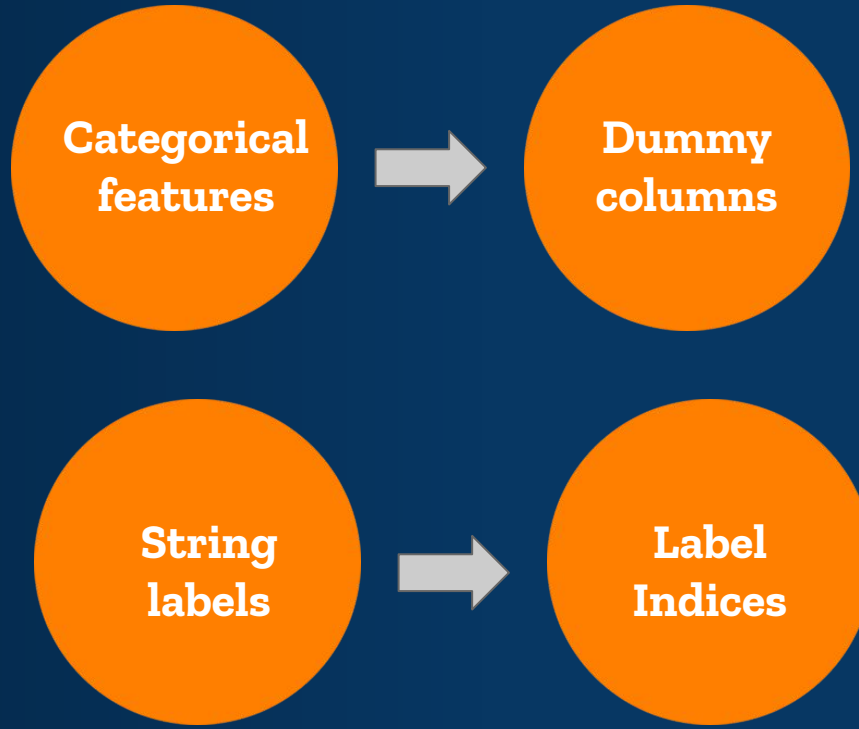
PIPELINE

Lecture Notes 10: ML Feature Utilities

Lecture Notes 11: ML Pipelines

- Apply StringIndexer
- Apply OneHotEncoder
- Apply VectorAssembler

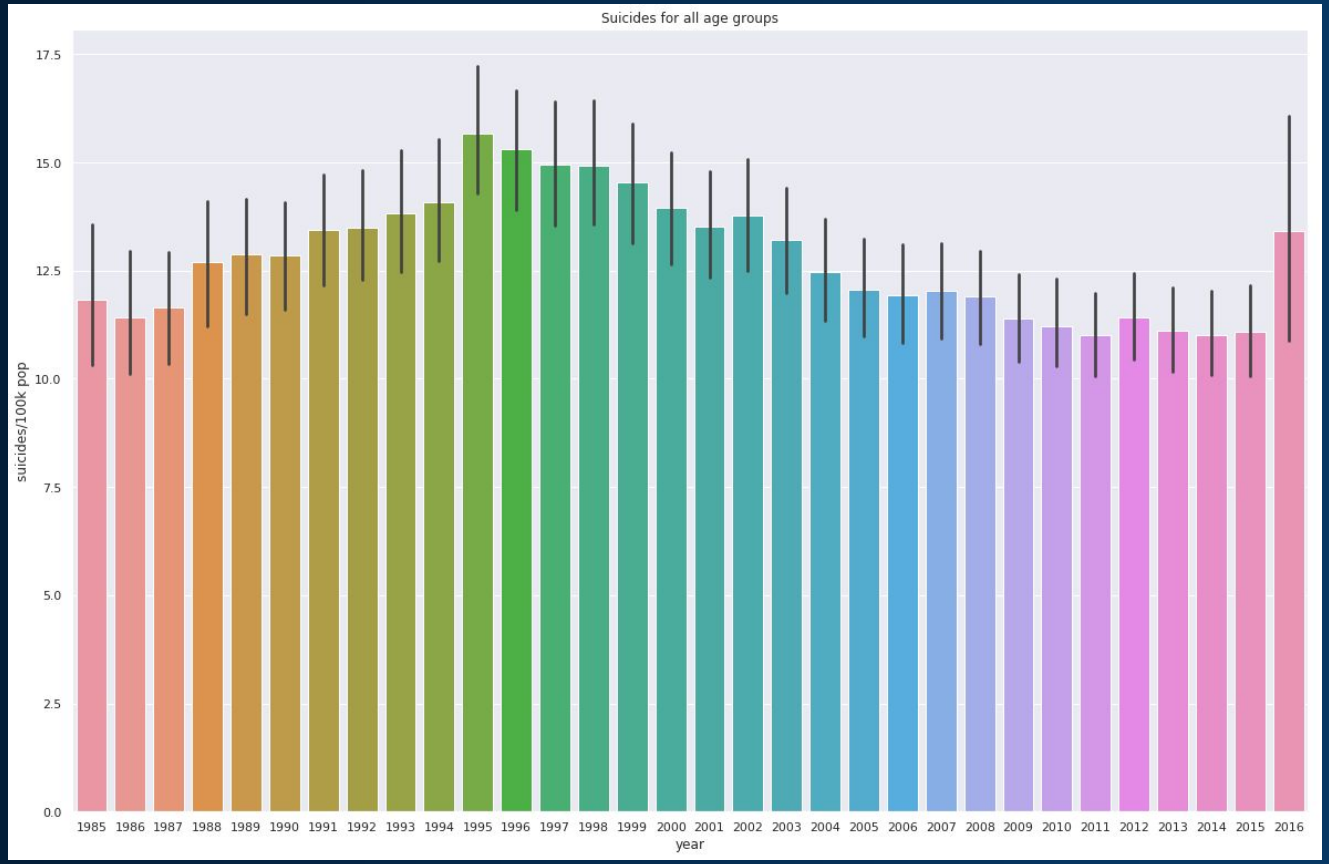
Model Construction



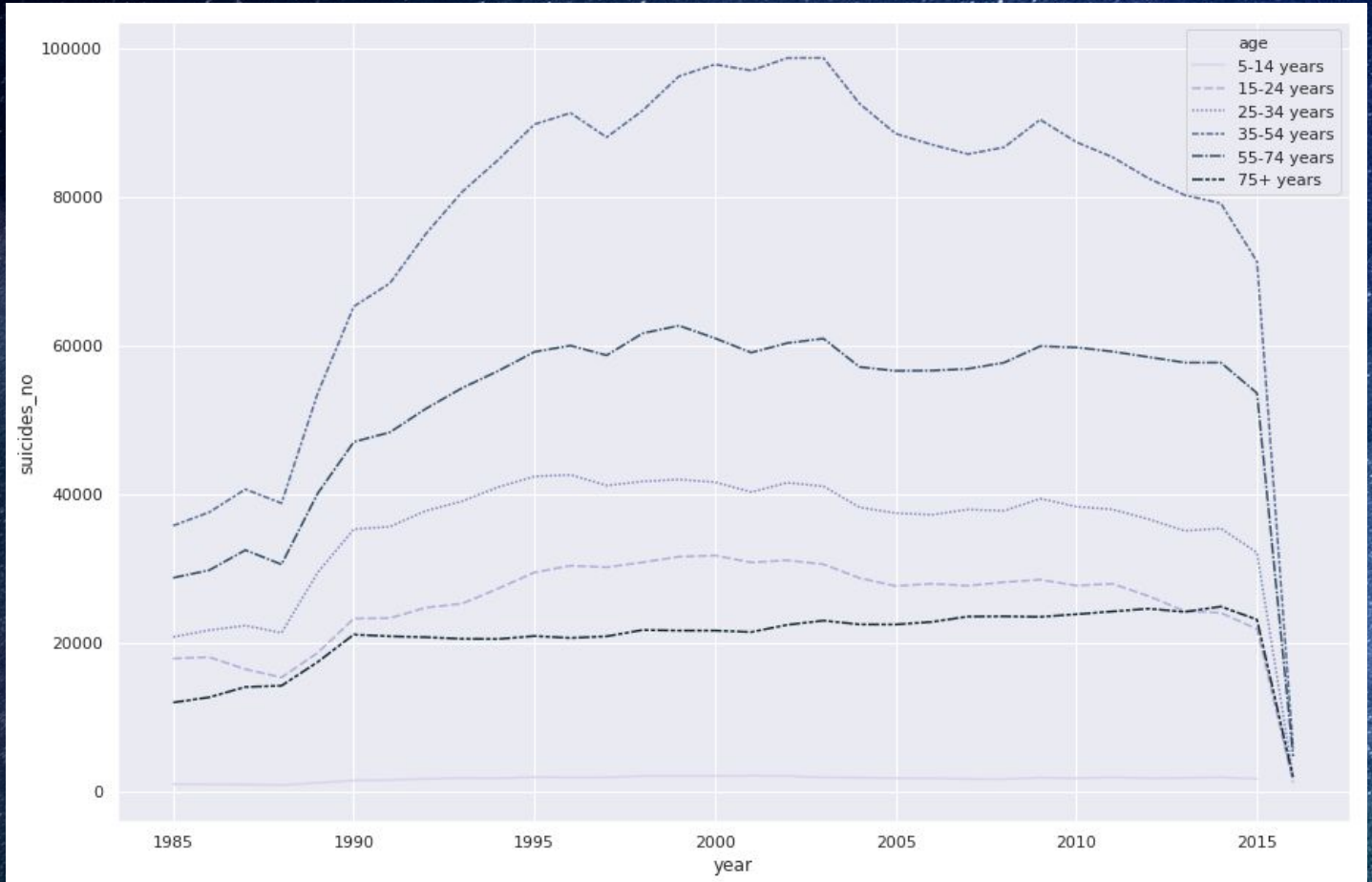
4. Exploratory Data Analysis

We mainly focus on the relation between Suicide Rate/ 100k and other factors.

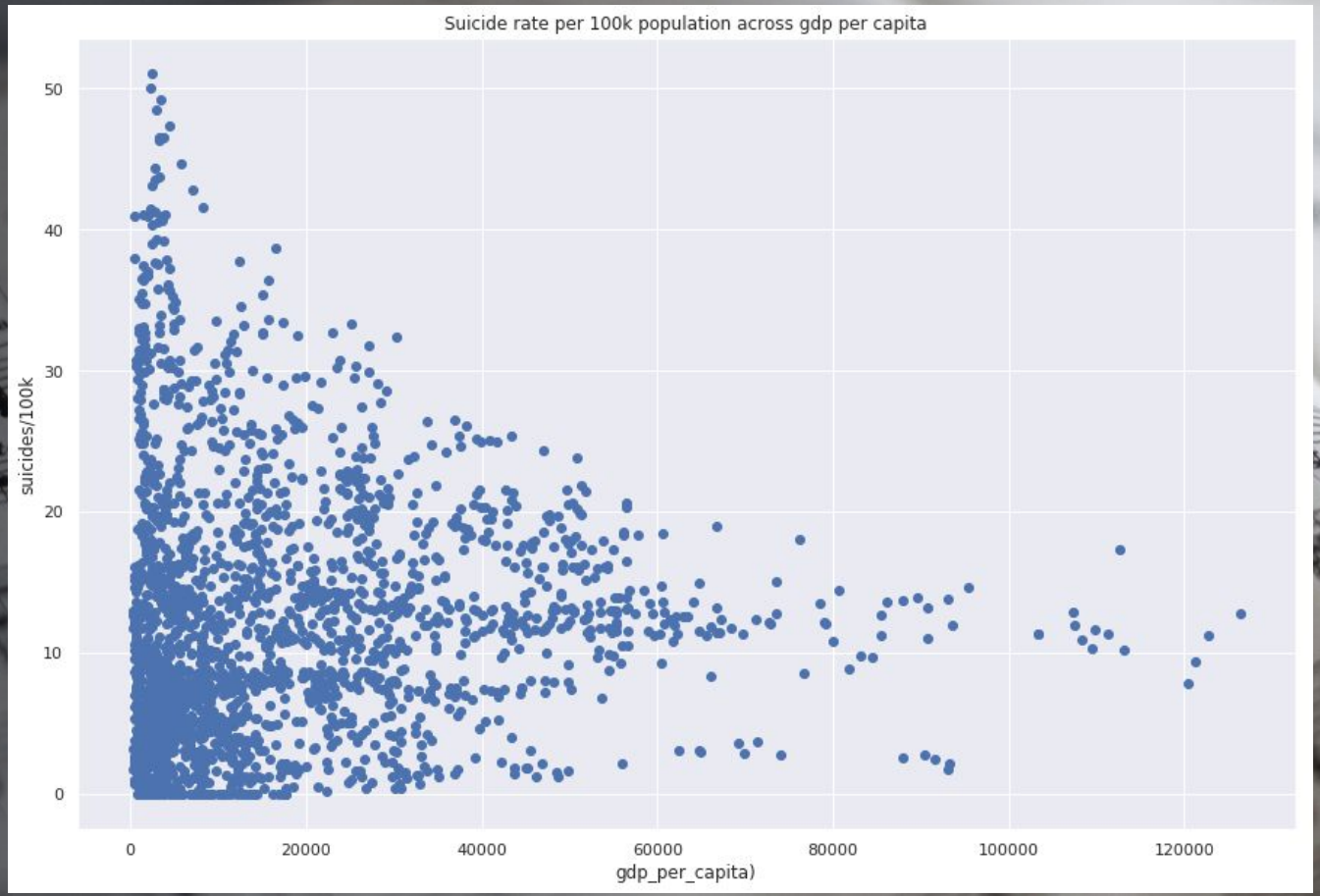
Visualization



Visualization



Visualization



5. Model Construction

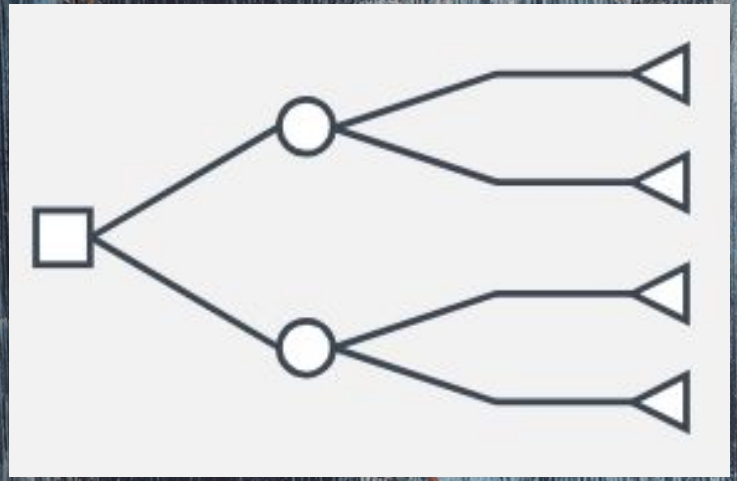
- Decision Tree
- Random Forest
- Gradient-Boosted Trees
- Linear Regression

Tree Methods: Decision Tree

Pros: non-parametric, easily interpretable, works well with incomplete and noisy data, can be used with regression and classification

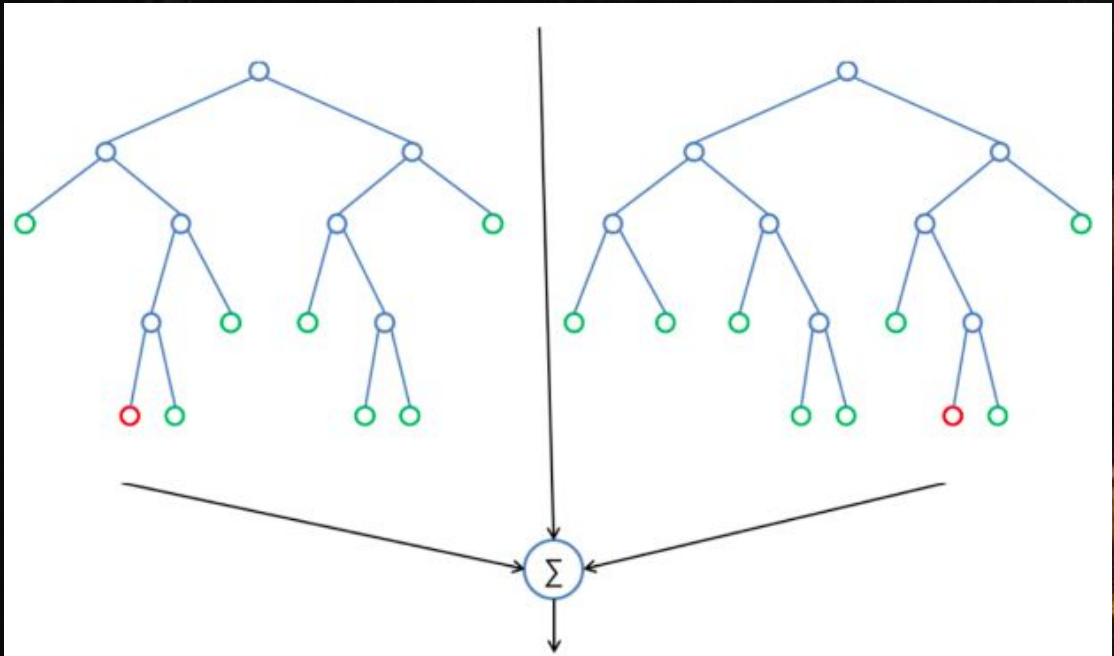
Cons: unstable, tends to overfit (low bias, high variance)

Lecture Notes 12: MLlib Regression



```
# Train a DecisionTree model.
dt = DecisionTreeRegressor(labelCol="label", featuresCol="features")
```

Tree Methods: Random Forest

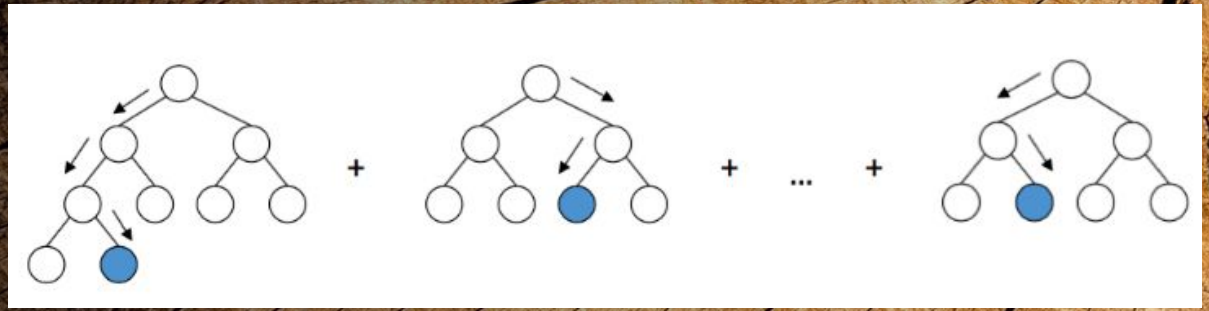



```
# Train a RandomForest model.
rf = RandomForestRegressor(labelCol="label", featuresCol="features")
```

Pros: lower classification error and better F-score than DTs, less likely to overfit (reduced variance), decorrelates trees, runs efficiently on large databases

Cons: not as easily interpretable, can overfit

Tree Methods: Gradient-Boosted Trees

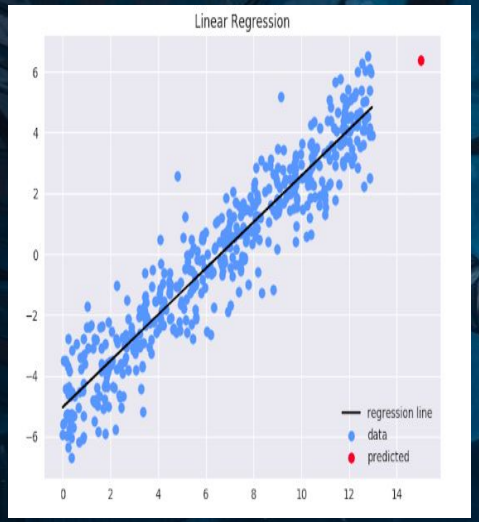


```
# Train a GBT model.
gbt = GBTRegressor(labelCol="label", featuresCol="features", maxIter=10)
```

Pros: model is more expressive (builds trees one at a time), performs better than RFs

Cons: training takes longer, harder to tune hyperparameters, more prone to overfitting, harder to get right

Linear Regression



Pros: simple, easy to explain, fast, no tuning required

Cons: can only use on numerical data, assumes linearity, outliers have large impact, missing values cannot be accommodated

Lecture Notes 12: MLlib Regression

```
# Train a Linear Regression model.
lr = LinearRegression(featuresCol = 'features', labelCol='label', maxIter=10, regParam=0.3, elasticNetParam=0.8)
```

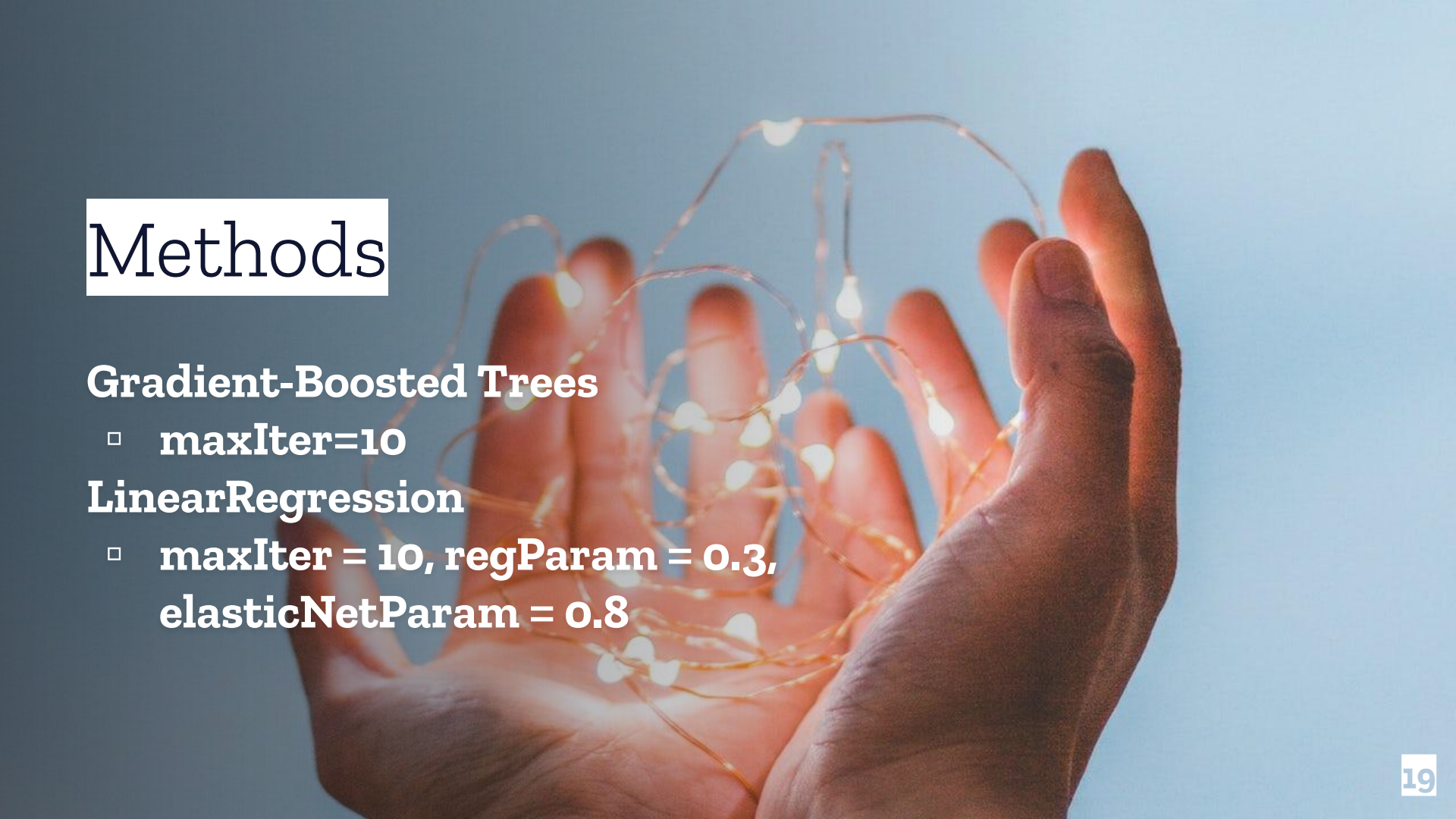
Methods

Gradient-Boosted Trees

- **maxIter=10**

LinearRegression

- **maxIter = 10, regParam = 0.3, elasticNetParam = 0.8**



6. Model Evaluation

Regression model evaluation

Regression analysis is used when predicting a continuous output variable from a number of independent variables.

Available metrics

Metric	Definition
Mean Squared Error (MSE)	$MSE = \frac{\sum_{i=0}^{N-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{N}$
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{\sum_{i=0}^{N-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{N}}$
Mean Absolute Error (MAE)	$MAE = \sum_{i=0}^{N-1} \lvert \mathbf{y}_i - \hat{\mathbf{y}}_i \rvert$
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{MSE}{\mathrm{VAR}(\mathbf{y}) \cdot (N-1)} = 1 - \frac{\sum_{i=0}^{N-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_{i=0}^{N-1}(\mathbf{y}_i - \bar{\mathbf{y}})^2}$
Explained Variance	$1 - \frac{\mathrm{VAR}(\mathbf{y} - \hat{\mathbf{y}})}{\mathrm{VAR}(\mathbf{y})}$

Which method should we employ?

Let's check out the RMSE values for each method

RMSE

Decision Tree	Random Forest	Gradient-Boosted Trees	Linear Regression
15.0669	15.0714	14.5837	15.7180

Best model of the 4: Gradient-Boosted Trees

We use this model to make our predictions

7. Conclusion and Future Studies

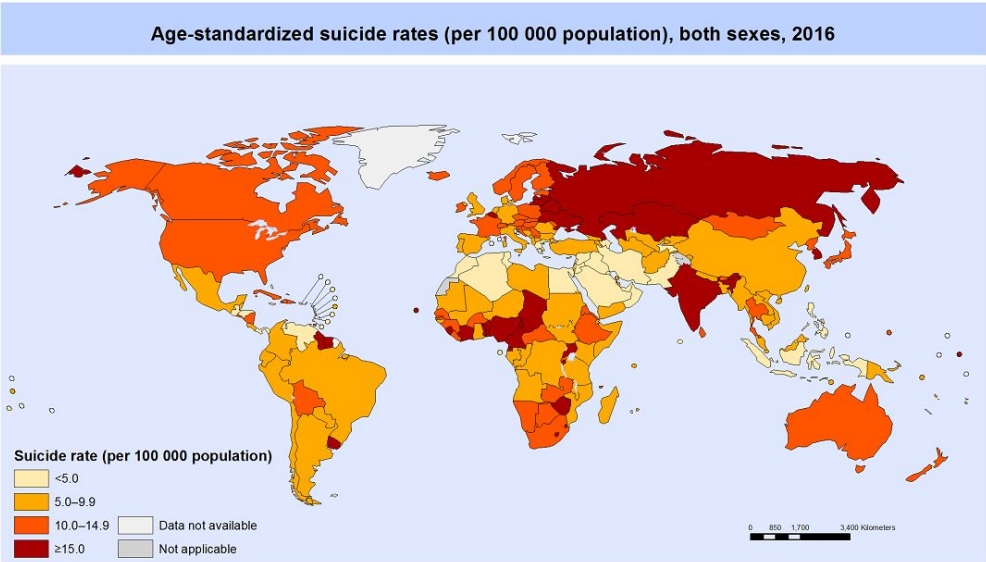
Future Studies

- K-Fold CV
- WHO to focus on countries and age groups with predicted higher suicide rates
- year column: time series
- Interesting topic: analyze text of clinical notes

RESEARCH ARTICLE

Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes

Suicide is a permanent solution to a temporary problem. -Phil Donahue




Thanks!