Predicting Personality and Traits Using Machine Learning

Senior Capstone I IT415 - 01

Dustin Powell

Cumberland University

Spring 2020 2/6/20

Introduction

- Personal data about people can be acquired from many sources.
- Much of this information is available through the internet.
  - Example: Social Media Post, Consumers Habits, etc.
- Information obtained can be used to predict other characteristics of an individual.
- The purpose of this work most accurately predict a person's personality and traits from this data using machine learning.

Case Study in PNAS

- Post–liking habits of roughly 58,500 people were able to predict various private aspects about the user such as age, race, intelligence, and political views [1].
- The primary tools were used in this research were linear and logistic regression [1].
- The users where assembled into a matrix called the User–Like matrix which shows the post that a user has liked relating to various things like products or hobbies [1].
- Singular Value Decomposition was then used on the set of data and then assembled into a User–Components Matrix to use the linear and logistic regression to make the predictions [1].
- The study has raised numerous ethical questions as well such as what would happen if the wrong person found out this information [1].

Questions to Address

The following questions arise from the case study by Kosinski, Stillwell, and Graepel [1]

Question 1

Using a machine learning algorithm what is the minimum amount of information needed to predict the traits and personality of a given person accurately?

Question 2

Can a person protect their private information about themselves?

Relationship Between Big Data and Predictive Analytics

- The previous case study is a specific example of a much larger and increasely important field.
- Big Data is the concept of collecting information from various online sources to be processed and or stored in databases to be used [3].
- Three types of data that are collected [3]:.
    - Structured
        - Example: Sales Figures
    - Unstructured
        - Example: Social Media Posts
    - Semi–Structured
        - Example: SQL Scripts, Server Logs
- Unstructured and semi–structured data have to be processed before storage versus structured that can be immediately stored into a database [3].

Forms of Analytics

- Three forms of analytics that are performed on the data stored in databases [3].
    - Descriptive Analytics
        - Analysis of data past occurrences of events.
    - Predictive Analytics
        - Analysis of data to make predictions.
    - Prescriptive Analytics
        - Analysis of data to make decisions.
- The application of statistics and computer science allow for these various forms of analytics to be conducted [3].

Overview of The Relationship

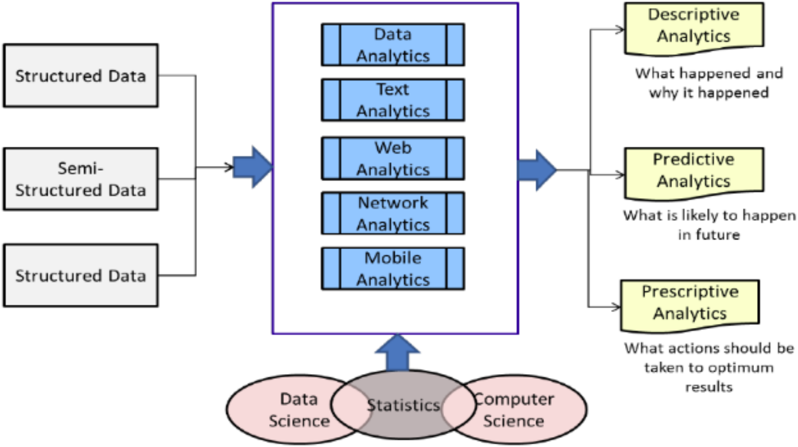

Figure: Visual representation, similar to a SQL ER–diagram, of the relationship between analytics and Big Data from the 2016 publication by Jeble, Kumari, and Patel [3].

Use of Predictive Analytics in Business

- Analytics is used by businesses to make educated decisions by using the resource as a way to understand markets [3].
- Predictive analytics can be used for predicting sales outcomes of a company's product versus the market competition [3].
- Marketing uses predictive analytics to determine which ads to deliver to internet users [3].
    - Example: Youtube Video Advertisements

Regression and Predictive Analytics

- Linear and multiple regression in general determines possible correlations between independent and dependent variables assuming a functional relationship between them to make predictions [2].
- In most cases multiple independent variables exist so multiple regression is used.
- Limitations in this technique are that functions must be assumed or that too many independent variables exist.

What is Machine Learning

- Machine Learning is a subsection of artificial intelligence that is specifically focused on using algorithms to make conclusions about a set of information [4].
- Three types of machine learning:
- Supervised Machine Learning
    - Using a base set of data to train the algorithm in which then makes predictions based on the learned information [4].
- Unsupervised Machine Learning
    - The use of an algorithm to classify and find relationships within a set of data without knowing of any results [4].
- Reinforcement Machine Learning
    - The use of a agent that learn from task carried out that results in a reward for correctly carrying out the task [4].

The Iris Data Set

- The iris data set is a classic data set of machine learning
- The data set is comprised of sepal and petal measurements of iris flowers[5].
- The set of information includes three types of iris called Setosa, Versicolor, and Virginica [5].
- Each of the flower has their sepal and petal width and length taken as the identify features of the flower [5].
- The data set was created by R. A. Fisher in 1936 for the purpose of pattern recognition [5].

Example of Mahcine Learning Algorithm Using Iris Data Set

```
# make predictions
from pandas import read_csv
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC

def main():
    # Load dataset
    csvData = "iris.csv"
    names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
    dataset = read_csv(csvData, names=names)
    # Split-out validation dataset
    array = dataset.values
    X = array[:,0:4]
    y = array[:,4]
    X_train, X_validation, Y_train, Y_validation = train_test_split(X, y, test_size=0.20, random_state=1)
    # Make predictions on validation dataset
    model = SVC(gamma='auto')
    model.fit(X_train, Y_train)
    predictions = model.predict(X_validation)
    # Evaluate predictions
    print(accuracy_score(Y_validation, predictions))
    print(confusion_matrix(Y_validation, predictions))
    print(classification_report(Y_validation, predictions))

main()
```

The algorithm is made by Brownlee to teach people about machine learning with slight edits to load to load a CSV file from the computer [6].

Sample Algorithm Output

Below is the sample output from running the algorithm:

```
0.966666666667
[[11  0  0]
 [ 0 12  1]
 [ 0  0  6]]
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00        11
Iris-versicolor       1.00      0.92      0.96        13
 Iris-virginica       0.86      1.00      0.92         6

    avg / total       0.97      0.97      0.97        30
```

Objectives to Accomplish

A brief overview of the upcoming objectives to accomplish.

1. Further defining and explanation of terminology and topics of machine learning
2. Locate a data set(s) of people's traits
3. Design a program using python to predict a person's traits
4. Test to determine if correct results about a person can be made.
5. Answer Question Two: Privacy?
6. Explain and adding additional details as needed.

References

- M. Kosinski, D. Stillwell, and T. Graepel (2013), ”Private traits and attributes are predicatable from digital records of human behavior”, *PNAS* 110. 15. , 5802 - 5805. Google Scholar. Accessed January 19th 2020 URL:https://www.pnas.org/content/pnas/110/15/5802.full.pdf

- J. Bothe, J. L. Brudney, and K. J. Meier (2015), *Applied Statistics For Public and Nonprofit Administration*. 4th ed. Stamford, CT USA Cengage Learning.

- S. Jeble, S. Kumari, and Y. Patil (2016), ”Role of big data and predictive analytics”. *International Journal of Automation and Logistics*. 2. 307-331. 10.1504/IJAL.2016.10001272. Google Scholar. Accessed January 21st 2020 URL: https://www.researchgate.net/publication/309809606

- V. Mirjalili and S. Raschka (2017), *Python Machine Learning*. 2nd ed. Birmingham, UK Packt Publishing.

- ”Iris Data Set”, University of California Irvine Machine Learning Repository. Accessed February 5th 2020 URL: https://archive.ics.uci.edu/ml/datasets/iris

- J. Brownlee (2019), ”Your First Machine Leaning Project in Python Step–By–Step”, Machine Learning Mastery. Accessed February 5ht 2020. URL: https://machinelearningmastery.com/machine-learning-in-python-step-by-step/