JUNE 1, 2020

Machine Learning Project for Cohort 5

AI SATURDAYS LAGOS

TABLE OF CONTENTS

BICYCLE SHARING DEMAND

Bicycle sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bicycle-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bicycle sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

For this problem, your group is expected to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program. After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for forecasting.

DATA FIELDS:

datetime	hourly date + timestamp
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	temperature in Celsius
atemp	"feels like" temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
count	number of total rentals

DATA SET: Bike Sharing

BREAST CANCER ANALYSIS AND PREDICTION

Breast cancer is the most common invasive cancer in women and the second leading cause of cancer death in women after lung cancer. Advances in screening and treatment for breast cancer have improved survival rates dramatically since 1989. Early screening, leading to detection has, from statistics, led to the increase in survival rates.

There is a possibility of detecting cancer at an early stage, based on features collected from the patients and your team has been tasked to research this possibility.

Features of the cell from various patients are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Analyse the data and build a model that would detect cancerous cells with good accuracy. After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for prediction.

DATA FIELDS:

ID	ID number
Diagnosis	M = malignant, B = benign
Ten real-valued features are computed for each cell nucleus	
radius	mean of distances from center to points on the perimeter
texture	standard deviation of gray-scale values
smoothness	local variation in radius lengths
perimeter	
area	
compactness	perimeter^2 / area - 1.0
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
symmetry	
fractal dimension	"coastline approximation" - 1

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.
Missing attribute values: none
Class distribution: 357 benign, 212 malignant

DATASET: Breast Cancer Wisconsin (Diagnostic) Data Set

CUSTOMER DEFECTION

Customer defection is the loss of clients or customers.

Telephone service companies, ISPs, insurance firms, etc, often use customer defection analysis and rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one.

Companies usually make a distinction between voluntary defection (churn) and involuntary defection (churn). Voluntary defection occurs due to a decision by the customer to switch to another company or service provider, involuntary defection occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary defection, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

Predictive analytics use defection prediction models that predict customer defection by assessing their propensity of risk to defection. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to defection.

Your team of machine learning engineers are presented with data from a telco company and tasked to analyse the data and gain insight into customer defection as-well-as develop a model that will predict the possibility of defection by any customer. After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for prediction.

DATASET: Customer Defection

RED WINE QUALITY

High quality red wine has been part of social, religious, and cultural events for hundreds of years. Medieval monasteries believed that their monks lived longer partly because of their regular, moderate drinking of high quality wine.

A restaurant in Nigeria is currently faced with the issue of quickly distinguishing wine by quality. They have reached out to your team to come up with a model that would help them quickly identify high quality wine.

You were able to access red wine data of Portuguese "Vinho Verde" wine through Cortez et al., 2009 research.

In this analysis, you will determine which physicochemical properties make red wine 'good!' by using some machine learning techniques. After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for prediction.

Tips: These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

Quality > 6.5 = "good"

DATASET: Red Wine Quality

LOAN PREDICTION

In finance, a loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations etc. The recipient (i.e., the borrower) incurs a debt and is usually liable to pay interest on that debt until it is repaid as well as to repay the principal amount borrowed.

XYZ is the world's largest online marketplace connecting borrowers and investors. An inevitable outcome of lending is default by borrowers. But usually, the company is faced with the problem of correctly identifying those who will default before lending.

They have called upon your team to help them develop a model that would correctly predict those who will default and result in a loan Charge Off.

In order to accomplish this, they have provided your team with their dataset that has information on those who received loans and those denied.

After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for prediction.

DATASET: Loan Prediction

CUSTOMER SEGMENTATION

Mr Ken owns a supermarket mall and through membership cards , he has some basic data about his customers like Customer ID, age, gender, annual income and spending score.

Spending Score is something he assigns to the customer based on defined parameters like customer behavior and purchasing data.

Mr Ken wants to understand the customers who can easily converge [Target Customers] so that the sense can be given to the marketing team and plan the strategy accordingly. He has reached out to your team for help.

After successful completion of the project, you are expected to prepare a report detailing your data analysis and information of algorithms used to develop models for segmentation.

DATASET: Customer Segmentation