greatlearning

Learning for Life

PREDICTIVE MODELING- WEEK-2

greatlearning
Learning for Life

DSBA CURRICULUM DESIGN

FOUNDATIONS

- Data Science Using Python
- Statistical Methods for Decision Making

CORE COURSES

- Advanced Statistics
- Data Mining
- Predictive Modelling(Week-2/ 5)
- Machine Learning
- Time Series Forecasting
- Data Visualization

DOMAIN APPLICATIONS

- Financial Risk Analytics
- Marketing Retail Analytics

greatlearning
Learning for Life

LEARNING OBJECTIVE OF THIS COURSE

- Linear Regression
- Logistic Regression
- Linear Discriminant Analysis

greatlearning
Learning for Life

LEARNING OBJECTIVES OF THIS SESSION

- Logistic Regression
- Sigmoid Function
- Understanding GridSearchCV for Logistic Regression

greatlearning
Learning for Life

TRY ANSWERING THE FOLLOWING

- Is Logistic regression a Supervised Machine Learning technique?
- How do we convert the log-odds output into the Probability output for Logistic Regression?
- Should we have a categorical dependent variable to apply Logistic Regression?

greatlearning
Learning for Life

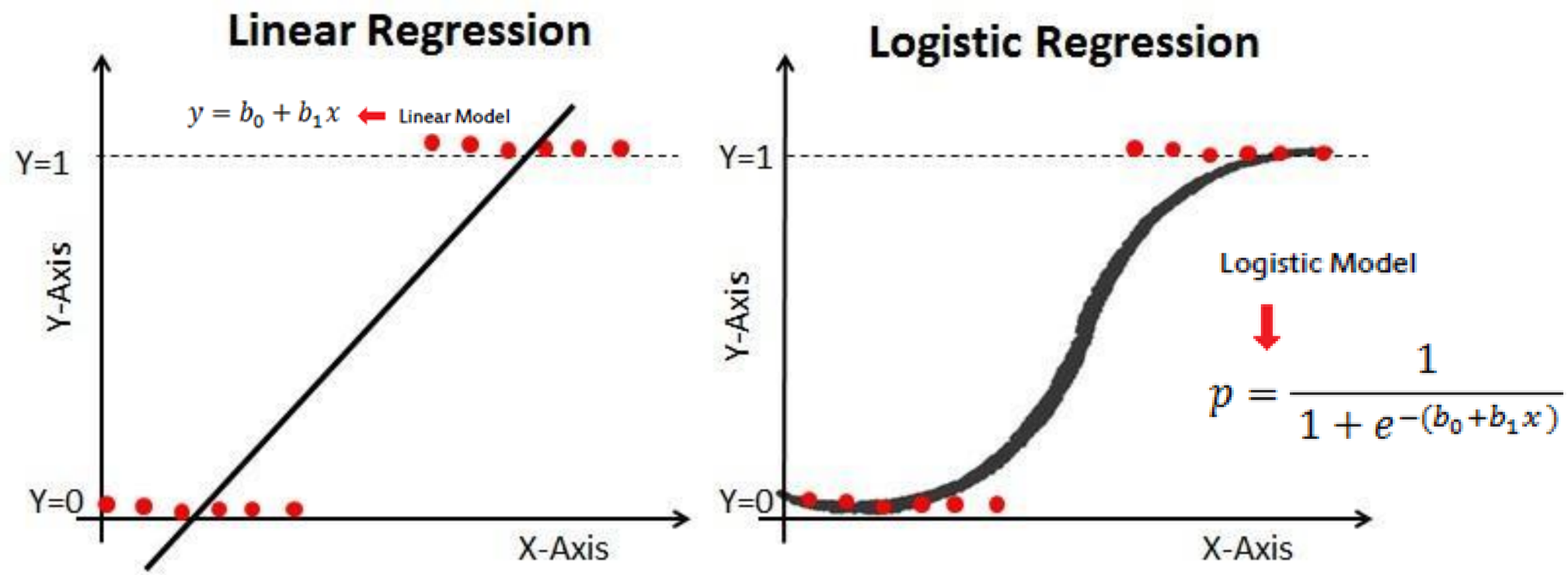
Why Logistic Regression?

Linear Regression helps us predicting continuous target variable but Logistic Regression helps us for predicting a discrete a target variable.
Logistic Regression is one of the “white-box” algorithms which helps us in determining the probability values and the corresponding cut-offs.

To understand more about the Logistic Regression function, please do visit the sklearn’s link over here.

greatlearning
Learning for Life

BROAD OVERVIEW



greatlearning
Learning for Life

Understanding GridSearchCV for Logistic Regression

GridSearchCV is a brute force iterative method of obtaining the best model based on a scoring metric provided by us and the parameters provided.

For all problems, accuracy might not be best possible metric based on which we can take a decision, so we need to be aware of other metrics such as precision, recall, f1-score etc.

To understand more about GridSearchCV please do visit the sklearn's link over here.

Industry Application - Logistic Regression for finding the best match

eHarmony is a popular online dating website. Nearly 4% of US marriages in 2012 were a result of eHarmony. According to Dr. Neil Clark Warren, many marriages ended in divorce because couples were not initially compatible. In 1997, Warren began an extensive research project interviewing 5000+ couples across the US, which became the basis of eHarmony's compatibility profile.

Therefore, unlike other online dating websites, eHarmony does not have users browse others' profiles. Instead, based on 29 different dimensions of personality including character, emotions, values, traits, etc assessed basis a detailed questionnaire, eHarmony computes a compatibility score between two people using logistic regression based predictive models and optimization techniques to determine their users' best matches.

Reference: https://www.cmo.com.au/article/635124/eharmony-how-machine-learning-leading-better-longer-lasting-love-matches/

Industry Application - Credit Risk Assessment using Logistic Regression

The Basel Committee is a representation of more than 60 Central Banks constituting more than 95% of world GDP. In its pursuit of monetary and financial stability, and to foster international cooperation it has rolled out Basel Accords. These Accords particularly address the areas of Risk such as Credit Risk, Operational Risk and Market Risk.

Logistic regression is by far the most widely accepted modeling approach in the area of Credit Risk both for retail(individual) and non-retail(institutional) exposures. While lending to an individual we use variables like age, income, education, employment status, emi to income ratio, and credit score etc. For corporate lending decisions the variables are mostly captured through the financial statements as they depict the true financial health of a company.

BANK FOR INTERNATIONAL SETTLEMENTS

Reference: https://www.listendata.com/2019/08/credit-risk-modelling.html

greatlearning
Learning for Life

CASE STUDY-

WHO is a specialized agency of the UN which is concerned with the world population health. Based upon the various parameters, WHO allocates budget for various areas to conduct various campaigns/initiatives to improve healthcare. Annual salary is an important variable which is considered to decide budget to be allocated for an area.

We have a data which contains information about 32561 samples and 15 continuous and categorical variables. Extraction of data was done from 1994 Census dataset.

The goal here is to build a binary model to predict whether the salary is >50K or <50K.

greatlearning
Learning for Life

ANY QUESTIONS

greatlearning
Learning for Life

HAPPY LEARNING