Predict Credit Card Default

Scenario

A credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. This would inform the issuer's decisions on who to give a credit card to and what credit limit to provide. It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers.

Data

The credit card issuer has gathered information on 30000 customers. The dataset contains information on 24 variables, including demographic factors, credit data, history of payment, and bill statements of credit card customers, as well as information on the outcome: did the customer default or not? The data to create the model is stored in the file "CreditCardDefault.csv".

Data Description:

Name	**Description**
ID	ID of each client
LIMIT_BAL	Amount of given credit in Dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-2=no consumption, -1=pay duly, 0=the use of revolving credit, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (Dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (Dollar)

BILL_AMT3	Amount of bill statement in July, 2005 (Dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (Dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (Dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (Dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (Dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (Dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (Dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (Dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (Dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (Dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

Modeling Customer Response

Scenario

A company that wants to achieve more profitable results by matching the right offer to each customer. To take advantage of the benefits of automation, a model is to be developed that predicts the reaction of a customer to an advertising campaign. Only these customers should then be contacted within a future campaign. In the past, data was collected for four campaigns, each targeted to a different type of customer account.

The goal is to predict the response to an offer.

Data

The data to create the model is stored in the file "CustomerResponse.csv". The file has historical data of 21,927 cases, each tracking the offer made to a specific customer in past campaigns, as indicated by the value of the campaign field. The coding is 1 = Standard account, 2 = Premium account, 3 = Gold account, and 4 = Platin account.

The file includes a response field that indicates whether the offer was accepted (0 = no, and 1 = yes). This will be the target field, that you want to predict.

A number of fields containing demographic and financial information about each customer are also included. There exist no detailed data description, since the responsible IT employee is no longer employed by the company.

Bank Marketing

A retail bank uses its own contact-center to do direct marketing campaigns, mainly through phone calls (telemarketing). Each campaign is managed in an integrated fashion and the results for all calls and clients within the campaign are gathered together, in a flat file report concerning only the data used to do the phone call.

Scenario

The objective of the project is to evaluate the efficiency and effectiveness of the telemarketing campaigns to sell long-term deposits. Therefore, on the one hand, the objective is to decrease the number of phone calls (efficiency dimension – cost reduction) and, on the other hand, to increase or at least not to decrease the total number of deposits subscriptions (effectiveness dimension – retain financial assets from clients for longer periods). In order to achieve this goal, the aim is to predict if the client will subscribe a term deposit.

Data

Data was collected mainly through the reports of previously executed campaigns. The data (BankMarketing.csv) contains the following 21 features organized in 41188 rows:

A. Bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')

6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')

7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

B. Related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular', 'telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

C. Other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

D. Social and economic context attributes

16 - emp.var.rate: employment variation rate, quarterly indicator (numeric)

17 - cons.price.idx: consumer price index, monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index, monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate, daily indicator (numeric)

20 - nr.employed: number of employees, quarterly indicator (numeric)

Output variable (desired target):

21 - y: has the client subscribed a term deposit? (binary: 'yes', 'no')

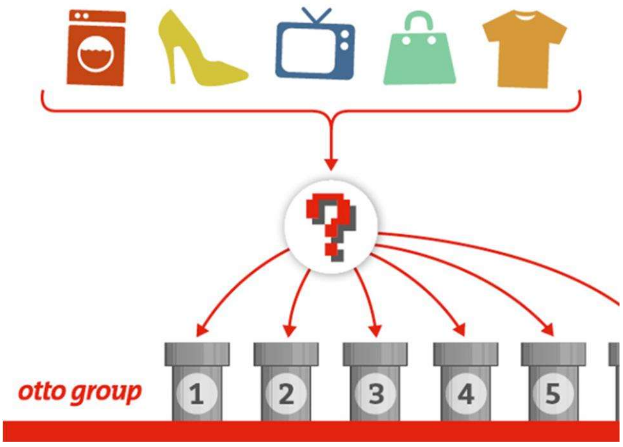
Classify products into the correct category

The Otto Group is one of the world's biggest e-commerce companies, with subsidiaries in more than 20 countries, including Crate & Barrel (USA), Otto.de (Germany) and 3 Suisses (France). They are selling millions of products worldwide every day, with always new products being added to the product line.

A consistent analysis of the performance of the products is crucial. However, due to the diverse global infrastructure, many identical products get classified differently. Therefore, the quality of the product analysis depends heavily on the ability to accurately cluster similar products. The better the classification, the more insights Otto can generate about the product range.

Scenario

The objective is to build a predictive model which is able to classify products into the correct category. There are nine categories for all products. Each target category represents one of the most important product categories (like fashion, electronics, etc.).



Data

For the task a dataset is provided in the form of a csv file (ClassifyProducts.csv) with 93 features for more than 200,000 products. The data contains the following features:

- id - an anonymous id unique to a product
- feat_1, feat_2, ..., feat_93 - the various features of a product
- target - the class of a product

Predict Potential Customers

This project is about a direct marketing case from the insurance sector which was to predict policy ownership. It is about predicting who would be interested in buying a caravan insurance policy.

Scenario

Direct mailings to a company's potential customers - "junk mail" to many - can be a very effective way for them to market a product or a service. However, as we all know, much of this junk mail is really of no interest to the people that receive it. Most of it ends up thrown away, wasting the money that the company spent on it.

If the company had a better understanding of who their potential customers were, they would know more accurately who to send it to, so some of this waste and expense could be reduced. The objective of this project is:

- Predict which customers are potentially interested in a caravan insurance policy.
- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

Data

The data to create the model is stored in the file "PredictPotentialCustomers.csv". It consists of 5822 real customer records. Each customer record consists of 86 variables, containing sociodemographic data (variables 1-43) and product ownership data (variables 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Variable 86 (Purchase), "CARAVAN: Number of mobile home policies", is the target variable which indicates whether the customer purchase a caravan insurance policy or not.

Data Description:

Nr. Name Description Domain

1 MOSTYPE Customer Subtype see L0

2 MAANTHUI Number of houses 1 – 10

3 MGEMOMV Avg size household 1 – 6

4 MGEMLEEF Avg age see L1

5 MOSHOOFD Customer main type see L2

6 MGODRK Roman catholic see L3

7 MGODPR Protestant ...

8 MGODOV Other religion

9 MGODGE No religion

10 MRELGE Married

11 MRELSA Living together

12 MRELOV Other relation

13 MFALLEEN Singles

14 MFGEKIND Household without children

15 MFWEKIND Household with children

16 MOPLHOOG High level education

17 MOPLMIDD Medium level education

18 MOPLLAAG Lower level education

19 MBERHOOG High status

20 MBERZELF Entrepreneur

21 MBERBOER Farmer

22 MBERMIDD Middle management

23 MBERARBG Skilled labourers

24 MBERARBO Unskilled labourers

25 MSKA Social class A

26 MSKB1 Social class B1

27 MSKB2 Social class B2

28 MSKC Social class C

29 MSKD Social class D

30 MHHUUR Rented house

31 MHKOOP Home owners

32 MAUT1 1 car

33 MAUT2 2 cars

34 MAUT0 No car

35 MZFONDS National Health Service

36 MZPART Private health insurance

37 MINKM30 Income < 30.000

38 MINK3045 Income 30-45.000

39 MINK4575 Income 45-75.000

40 MINK7512 Income 75-122.000

41 MINK123M Income >123.000

42 MINKGEM Average income

43 MKOOPKLA Purchasing power class

44 PWAPART Contribution private third party insurance see L4

45 PWABEDR Contribution third party insurance (firms) ...

46 PWALAND Contribution third party insurane (agriculture)

47 PPERSAUT Contribution car policies

48 PBESAUT Contribution delivery van policies

49 PMOTSCO Contribution motorcycle/scooter policies

50 PVRAAUT Contribution lorry policies

51 PAANHANG Contribution trailer policies

52 PTRACTOR Contribution tractor policies

53 PWERKT Contribution agricultural machines policies

54 PBROM Contribution moped policies

55 PLEVEN Contribution life insurances

56 PPERSONG Contribution private accident insurance policies

57 PGEZONG Contribution family accidents insurance policies

58 PWAOREG Contribution disability insurance policies

59 PBRAND Contribution fire policies

60 PZEILPL Contribution surfboard policies

61 PPLEZIER Contribution boat policies

62 PFIETS Contribution bicycle policies

63 PINBOED Contribution property insurance policies

64 PBYSTAND Contribution social security insurance policies

65 AWAPART Number of private third party insurance 1 - 12

66 AWABEDR Number of third party insurance (firms) ...

67 AWALAND Number of third party insurane (agriculture)

68 APERSAUT Number of car policies

69 ABESAUT Number of delivery van policies

70 AMOTSCO Number of motorcycle/scooter policies

71 AVRAAUT Number of lorry policies

72 AAANHANG Number of trailer policies

73 ATRACTOR Number of tractor policies

74 AWERKT Number of agricultural machines policies

75 ABROM Number of moped policies

76 ALEVEN Number of life insurances

77 APERSONG Number of private accident insurance policies

78 AGEZONG Number of family accidents insurance policies

79 AWAOREG Number of disability insurance policies

80 ABRAND Number of fire policies

81 AZEILPL Number of surfboard policies

82 APLEZIER Number of boat policies

83 AFIETS Number of bicycle policies

84 AINBOED Number of property insurance policies

85 ABYSTAND Number of social security insurance policies

86 CARAVAN Number of mobile home policies 0 - 1

L0:

Value Label

1 High Income, expensive child

2 Very Important Provincials

3 High status seniors

4 Affluent senior apartments

5 Mixed seniors

6 Career and childcare

7 Dinki's (double income no kids)

8 Middle class families

9 Modern, complete families

10 Stable family

11 Family starters

12 Affluent young families

13 Young all american family

14 Junior cosmopolitan

15 Senior cosmopolitans

16 Students in apartments

17 Fresh masters in the city

18 Single youth

19 Suburban youth

20 Etnically diverse

21 Young urban have-nots

22 Mixed apartment dwellers

23 Young and rising

24 Young, low educated

25 Young seniors in the city

26 Own home elderly

27 Seniors in apartments

28 Residential elderly

29 Porchless seniors: no front yard

30 Religious elderly singles

31 Low income catholics

32 Mixed seniors

33 Lower class large families

34 Large family, employed child

35 Village families

36 Couples with teens 'Married with children'

37 Mixed small town dwellers

38 Traditional families

39 Large religous families

40 Large family farms

41 Mixed rurals

L1:

1 20-30 years

2 30-40 years

3 40-50 years

4 50-60 years

5 60-70 years

6 70-80 years

L2:

1 Successful hedonists

2 Driven Growers

3 Average Family

4 Career Loners

5 Living well

6 Cruising Seniors

7 Retired and Religeous

8 Family with grown ups

9 Conservative families

10 Farmers

L3:

0 0%

1 1 - 10%

2 11 - 23%

3 24 - 36%

4 37 - 49%

5 50 - 62%

6 63 - 75%

7 76 - 88%

8 89 - 99%

9 100%

L4:

0 f 0

1 f 1 – 49

2 f 50 – 99

3 f 100 – 199

4 f 200 – 499

5 f 500 – 999

6 f 1000 – 4999

7 f 5000 – 9999

8 f 10.000 - 19.999

9 f 20.000 - ?

Prediction of returns

Returns constitute a very significant cost factor for online retailers. The resulting costs have to be borne by the trader. Especially in the clothing trade returns proportions of partially more than 50% are not exceptional. The goal for the sender is to lower these proportions without causing deterioration in customer service. It becomes evident that preventive measures carried out on the basis of probabilities of returns (restriction with respect to payment options, adjustment of shipping costs, sizing guides, …) could become a target-oriented strategy.

Scenario

On the basis of historical purchase data of an online shop a model is to be created for predicting if a certain purchase is converted into a return or not. For this purpose the historical data contain as well purchase and shipping data as different product and customer attributes. The information “return yes/no” is the target variable.

Data

For the task anonymized real shop data are provided in the form of a csv file consisting of individual data sets. The data (ReturnsPrediction.csv) contains the following features:

Column name	Description	Range of values	Existence of missing values
orderItemID	Number of the order item	Natural number	No
orderDate	Order date	Date	No
deliveryDate	Delivery date	Date	Yes
itemID	Item ID	Natural number	No
size	Size of the item	String	No
color	Color of the time	String	Yes
manufacturerID	Manufacturer ID	Natural number	No
price	Price of the item	Positive real number	No
customerID	Customer ID	Natural number	No
salutation	Salutation of the customer	String	No
dateOfBirth	Customer’s date of birth	Date	Yes
state	Federal state of the customer	String	No
creationDate	Date of account creation	Date	No
returnShipment	Return no/yes	{0, 1}	No

Predicting Avocado Prices

The Hass avocado is a cultivar of avocado with dark green-colored, bumpy skin. The price of a single avocado depends on different factors.

Scenario

Your task is to predict the prices according to the factors.

Data

There are 13 attributes in each case of the dataset (AvocadoPrices.csv). They are:

- Date - The date of the observation
- AveragePrice - the average price of a single avocado
- Total Volume - Total number of avocados sold
- 4046 - Total number of avocados with PLU 4046 sold
- 4225 - Total number of avocados with PLU 4225 sold
- 4770 - Total number of avocados with PLU 4770 sold
- Total Bags, Small Bags, Large Bags, XLarge Bags – number of bags of specific type
- type - conventional or organic
- year - the year
- Region - the city or region of the observation

Avocados are normally stickered with a PLU number which will help you identify what kind and where it’s from. PLU numbers are used to properly identify the size of avocados sold at retail. The numerical column names refer to price lookup codes:

- 4046: small Hass
- 4225: large Hass
- 4770: extra large Hass

Bike Sharing Rental Demand Estimation

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, users are able to easily rent a bike from a particular position and return back at another position. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing systems into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

Scenario

The bike sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The objective is to build a predictive model which is able to predict the number of bike rentals within a specific hour based on the environmental and seasonal settings.

Data

The Bike Rental UCI dataset (BikeRental.csv) contains 17,379 rows and 17 columns, each row representing the number of bike rentals within a specific hour of a day in the years 2011 or 2012. Weather conditions (such as temperature, humidity, and wind speed) are included in this feature set, and the dates are categorized as holiday vs. weekday etc. The field to predict is "cnt", which contain a count value ranging from 1 to 977, representing the number of bike rentals within a specific hour.

The data contains the following features:

- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : if day is holiday or not
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
  - 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 : Normalized temperature in Celsius.
- atemp: Normalized feeling temperature in Celsius.
- hum: Normalized humidity.
- windspeed: Normalized wind speed.
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Big Mart Sales Prediction

Big Mart is an online one stop marketplace where you can buy or sell or advertise your merchandise at low cost. Their goal is to make Big Mart the shopping paradise for buyers and the marketing solution for sellers. The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined.

Scenario

The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Data

There are 12 attributes in each of the 8523 cases of the dataset (BigMart.csv). They are:

- Item_Identifier - Unique product ID
- Item_Weight - Weight of product
- Item_Fat_Content - Whether the product is low fat or not
- Item_Visibility - The % of total display area of all products in a store allocated to the particular product
- Item_Type - The category to which the product belongs
- Item_MRP - Maximum Retail Price (list price) of the product
- Outlet_Identifier - Unique store ID
- Outlet_Establishment_Year - The year in which store was established
- Outlet_Size - The size of the store in terms of ground area covered
- Outlet_Location_Type - The type of city in which the store is located
- Outlet_Type - Whether the outlet is just a grocery store or some sort of supermarket
- Item_Outlet_Sales - Sales of the product in the particular store [target variable to be predicted].

Retail Company Sales Forecasting

Predicting future sales is one of the most important aspects of strategic planning for a retail company. The retail company "We Sell Everything" has 45 stores across the country. They are interested in predicting sales volumes for the different stores

The data was collected from 2010 to 2012. It provides information on the historical sales data of the 45 stores. Additionally, external data is available (like CPI, Unemployment Rate and Fuel Prices in the region of each store) which might help to make a more detailed analysis.

The retail company is also known for conducting promotional markdown events before major holidays such as Christmas and Easter among others. The difference between the weightage given to the data of regular weeks and the weeks including holiday seasons, is of additional interest.

Scenario

The main goal is to predict the sales of each store. Furthermore, the effects of markdowns on the sales during the holiday seasons should be analyzed and predicted.

Data

The dataset includes 421,570 observations. There are 20 features in the dataset (StoresData.csv). They are:

Feature	Definition
CPI	Consumer Price Index during that week.
Date	The date where this observation was taken.
Fuel_Price	Fuel Price in that region during that week.
IsHoliday	Boolean value representing a holiday week or not.
MarkDown1-5	Represents the Type of markdown and what quantity was available during that week.
Size	Sets the size of a Store would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.
Store	The store number. Range from 1–45.
Temperature	Temperature of the region during that week.
Type	Three types of stores ‘A’, ‘B’ or ‘C’.
Unemployment	The unemployment rate during that week in the region of the store.
Weekly_Sales	The sales recorded during that Week.

Month	Number of the month.
MonthMean	Mean of the month over all stores.
StoreMean	Mean of the store over all month.
Store_MonthMean	Mean of the month of the particular store.
Week	Number of the week.