# DAT 650 Use Cases Document

All of the accompanying data sets can be found in the **Assignment Guidelines and Rubrics** area of your course.

## Credit/Financing Risk

In response to the financial crisis of 2008–2009, the credit branch has been asked to reassess the method used to determine if an application presents a bad credit risk. The example file has data on 1,000 past credit applicants, described by 31 variables. The goal is to determine the likelihood of default for a new credit application based on a subset of the 31 variables. The business user will use this model to help determine the risk associated with extending the credit applicant credit, based on the data provided on the application. The impact of the dollar loss to the company on the default of a customer loan is 150% of the remaining balance.

The current data environment is an Excel spreadsheet that contains information about the active credit applicants assigned to a business user. The spreadsheet allows the business user to adjust the values of certain variables like loan amount to generate the scenario-based predictive likelihood that the customer will default. The model will run nightly for all loan applicants and current loan holders with reports generated for high-risk loan applicants/holders.

The data is stored in an Oracle database in transactional form. The IT department has built a data warehouse that is updated each evening with the current day data. The credit team uses this data warehouse data as the source for their reports and has the ability to have an ad hoc extract to select data into Excel for their unique research needs. The credit team extracts are limited in rows as well as fields that have been corporately pre-approved for extraction.

GE has compiled a file for the purpose of this pilot project using the extract tool. The dataset provided, Credit Data (XLS), includes information on credit applicants which the GE credit team believes to be relevant to analyzing this problem.

The credit team would like to determine if this data can be used to identify credit default of loan applicants. It is important to be able to understand loan default drivers for metadata like salary, interest rates, and other pertinent groupings which come from the analysis.

The pilot will need to only show basis for this data to be able to describe and generally identify credit applicants that may default. The management team expects to make a GO or NO GO business decision based on the pilot recommendation. If there is a GO, then GE will allocate new project dollars to arrange for GE resources to develop a full-enterprise deployed predictive analytic model. Note that the results of this pilot will be used as a basis for that next project.

## Data Description/Documentation for Credit/Financing Risk

These descriptions are also available in the Credit Data Descriptions (CSV) file in the Assignment Guidelines and Rubrics area of the course.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Var. #** | **Variable Name** | **Description** | **Variable Type** | **Description** |
| 1. | OBS# | Observation No. | Categorical |  |
| 2. | CHK\_ACCT | Checking account status | Categorical | 0 : < 0 |
|  |  |  |  | 1: 0 < ...< 200 |
|  |  |  |  | 2 : => 200 |
|  |  |  |  | 3: no checking account |
| 3. | DURATION | Duration of credit in months | Numerical |  |
| 4. | HISTORY | Credit history | Categorical | 0: no credits taken |
|  |  |  |  | 1: all credits at this bank paid back duly |
|  |  |  |  | 2: existing credits paid back duly till now |
|  |  |  |  | 3: delay in paying off in the past |
|  |  |  |  | 4: critical account |
| 5. | NEW\_CAR | Purpose of credit | Binary | car (new) 0: No, 1: Yes |
| 6. | USED\_CAR | Purpose of credit | Binary | car (used) 0: No, 1: Yes |
| 7. | FURNITURE | Purpose of credit | Binary | furniture/equipment 0: No, 1: Yes |
| 8. | RADIO/TV | Purpose of credit | Binary | radio/television 0: No, 1: Yes |
| 9. | EDUCATION | Purpose of credit | Binary | Education 0: No, 1: Yes |
| 10. | RETRAINING | Purpose of credit | Binary | Retraining 0: No, 1: Yes |
| 11. | AMOUNT | Credit amount | Numerical |  |
| 12. | SAV\_ACCT | Average balance in savings account | Categorical | 0 : < 100 |
|  |  |  |  | 1 : 100<= ... < 500 |
|  |  |  |  | 2 : 500<= ... < 1000 |
|  |  |  |  | 3 : =>1000 |
|  |  |  |  | 4 : unknown/ no savings account |
| 13. | EMPLOYMENT | Present employment since | Categorical | 0 : unemployed |
|  |  |  |  | 1: < 1 year |
|  |  |  |  | 2 : 1 <= ... < 4 years |
|  |  |  |  | 3 : 4 <=... < 7 years |
|  |  |  |  | 4 : >= 7 years |
| 14. | INSTALL\_RATE | Installment rate as % of disposable income | Numerical |  |
| 15. | MALE\_DIV | Applicant is male and divorced | Binary | 0: No, 1: Yes |
| 16. | MALE\_SINGLE | Applicant is male and single | Binary | 0: No, 1: Yes |
| 17. | MALE\_MAR\_WID | Applicant is male and married or a widower | Binary | 0: No, 1: Yes |
| 18. | CO-APPLICANT | Application has a co-applicant | Binary | 0: No, 1: Yes |
| 19. | GUARANTOR | Applicant has a guarantor | Binary | 0: No, 1: Yes |
| 20. | PRESENT\_RESIDENT | Present resident since - years | Categorical | 0: <= 1 year |
|  |  |  |  | 1<…<=2 years |
|  |  |  |  | 2<…<=3 years |
|  |  |  |  | 3:>4years |
| 21. | REAL\_ESTATE | Applicant owns real estate | Binary | 0: No, 1: Yes |
| 22. | PROP\_UNKN\_NONE | Applicant owns no property (or unknown) | Binary | 0: No, 1: Yes |
| 23. | AGE | Age in years | Numerical |  |
| 24. | OTHER\_INSTALL | Applicant has other installment plan credit | Binary | 0: No, 1: Yes |
| 25. | RENT | Applicant rents | Binary | 0: No, 1: Yes |
| 26. | OWN\_RES | Applicant owns residence | Binary | 0: No, 1: Yes |
| 27. | NUM\_CREDITS | Number of existing credits at this bank | Numerical |  |
| 28. | JOB | Nature of job | Categorical | 0 : unemployed/ unskilled - non-resident |
|  |  |  |  | 1 : unskilled - resident |
|  |  |  |  | 2 : skilled employee / official |
|  |  |  |  | 3 : management/ self-employed/highly qualified employee/ officer |
| 29. | NUM\_DEPENDENTS | Number of people for whom liable to provide maintenance | Numerical |  |
| 30. | TELEPHONE | Applicant has phone in his or her name | Binary | 0: No, 1: Yes |
| 31. | FOREIGN | Foreign worker | Binary | 0: No, 1: Yes |
| 32 | DEFAULT | Defaulted on loan | Binary | 0: No, 1: Yes |

## Customer Churn

*Churn* (a derivative of *change* and *turn*), also expressed as *customer attrition* or *customer turnover,* refers to the loss of customers or clients. Within the GE Healthcare portfolio, a Health IT service is available on a subscription basis. The service is used by smartphones and provides a platform to assist medical professionals in their daily roles. The service includes a required use of a GE-provided smartphone and cellular service configuration to ensure the data streaming bandwidth is available and sufficient to support the medical professionals’ use of the smartphone application. This GE cellular service provides the medical professional with all normal cell phone usage in addition to the application.

GE has partnered with a cellular carrier–SmartAppCellular–that provides dedicated bandwidth and configuration services for cell phone applications. GE is beginning to experience a small amount of attrition, and based on customer feedback, it is related to the cellular service and not the application. GE Healthcare recognizes that other vendors are beginning to compete in this space and is attempting to identify ways to retain its customers.

The current data environment is a database that contains subscriber information which is updated at the end of each month and reflects the most current month-end cumulative account information. The data is stored in an Oracle database in account-level transactional form. The IT department has built a data warehouse that is updated each end of month by appending the most recent end-of-month data.

The customer account management team has access to this information through monthly data warehouse reporting and has access to generate ad hoc extracts of select data into Excel for their unique research needs. The team extracts are limited in rows as well as fields that have been corporately pre-approved for extraction.

GE has compiled a file for the purpose of this pilot project using the extract tool. The dataset provided in the **Assignment Guidelines and Rubrics** area of the course, Cell Data (CSV file), includes information on GE Healthcare IT Application subscribers which the GE Customer Account Management Team believes to be relevant to analyzing this problem.

The Customer Account Management team would like to determine if this data can be used to identify subscribers that may churn. It is important to be able to understand churn default drivers for metadata like longevity, cell usage, and other pertinent groupings which come from the analysis.

The pilot will need to only show basis for this data to be able to describe and generally identify subscribers that may churn. The management team expects to make a GO or NO GO business decision based on the pilot recommendation. If there is a GO, then GE will allocate new project dollars to arrange for GE resources to develop a full-enterprise deployed predictive analytic model. Note that the results of this pilot will be used as a basis for that next project.

## Data Description/Documentation for Customer Churn

These descriptions are also available in the Cell Data Descriptions file (XLSX) in the Assignment Guidelines and Rubrics area of the course.

Note: Variables with description beginning with “Mean” are averages calculated over the four months previous to observing churn.

|  |  |  |  |
| --- | --- | --- | --- |
| **Position** | **Variable Name** | **Variable Description** | **Notes:** |
| 1 | REVENUE | Mean monthly revenue |  |
| 2 | MOU | Mean monthly minutes of use |  |
| 3 | RECCHRGE | Mean total recurring charge | The recurring charge is the basic rate for the customer's calling plan |
| 4 | DIRECTAS | Mean number of director assisted calls |  |
| 5 | OVERAGE | Mean overage minutes of use | Overage represents calls or minutes of use over the number of minutes allowed by that customer's calling plan |
| 6 | ROAM | Mean number of roaming calls |  |
| 7 | CHANGEM | % Change in minutes of use |  |
| 8 | CHANGER | % Change in revenues |  |
| 9 | DROPVCE | Mean number of dropped voice calls |  |
| 10 | BLCKVCE | Mean number of blocked voice calls |  |
| 11 | UNANSVCE | Mean number of unanswered voice calls |  |
| 12 | CUSTCARE | Mean number of customer care calls | Customer Care handles customer calls regarding complaints, questions, and so on |
| 13 | THREEWAY | Mean number of threeway calls |  |
| 14 | MOUREC | Mean unrounded mou received voice calls |  |
| 15 | OUTCALLS | Mean number of outbound voice calls |  |
| 16 | INCALLS | Mean number of inbound voice calls |  |
| 17 | PEAKVCE | Mean number of in and out peak voice calls |  |
| 18 | OPEAKVCE | Mean number of in and out off-peak voice calls |  |
| 19 | DROPBLK | Mean number of dropped or blocked calls |  |
| 20 | CALLFWDV | Mean number of call forwarding calls |  |
| 21 | CALLWAIT | Mean number of call waiting calls |  |
| 22 | CHURN | Churn between 31-60 days after obs\_date | 1=> the customer churned; 0=> the customer did not churn |
| 23 | MONTHS | Months in Service | Number of months the customer has had service. Number of individuals listed with the account |
| 24 | UNIQSUBS | Number of Uniq Subs |  |
| 25 | ACTVSUBS | Number of Active Subs | Number of individuals listed with the account who actively use the service. Location of the customer within the U.S. |
| 26 | CSA | Communications Service Area |  |
| 27 | PHONES | # Handsets Issued |  |
| 28 | MODELS | # Models Issued |  |
| 29 | EQPDAYS | Number of days of the current equipment |  |
| 30 | CUSTOMER | Customer ID |  |
| 31 | AGE1 | Age of first household member |  |
| 32 | AGE2 | Age of second household member |  |
| 33 | CHILDREN | Presence of children in household |  |
| 34 | CREDITA | Highest credit rating - a |  |
| 35 | CREDITAA | High credit rating - aa |  |
| 36 | CREDITB | Good credit rating - b |  |
| 37 | CREDITC | Medium credit rating - c |  |
| 38 | CREDITDE | Low credit rating - de |  |
| 39 | CREDITGY | Very low credit rating - gy |  |
| 40 | CREDITZ | Lowest credit rating - z |  |
| 41 | CREDIT\_RATING | Ranked credit rating | 1 = aa, 2 = a, 3 = b, 4 = c, 5 = de, 6 = gy, 7 = z |
| 42 | PRIZMRUR | Prizm code is rural | Prizm data classifies the customer's location by rural, suburban, or town |
| 43 | PRIZMUB | Prizm code is suburban |  |
| 44 | PRIZMTWN | Prizm code is town |  |
| 45 | PRZM\_NUM | Prizm code - unified | 0 = Unknown, 1 = Rural, 2 = Urban, 3 = Town |
| 46 | REFURB | Handset is refurbished |  |
| 47 | WEBCAP | Hanset is web capable |  |
| 48 | TRUCK | Subscriber owns a truck |  |
| 49 | RV | Subscriber owns a recreational vehicle |  |
| 50 | OCCPROF | Occupation - professional |  |
| 51 | OCCCLER | Occupation - clerical |  |
| 52 | OCCCRFT | Occupation - crafts |  |
| 53 | OCCSTUD | Occupation - student |  |
| 54 | OCCHMKR | Occupation - homemaker |  |
| 55 | OCCRET | Occupation - retired |  |
| 56 | OCCSELF | Occupation - self-employed |  |
| 57 | OCC | Occupation -Unified | 0 = Unknown, 1 = Professional, 2 = Clerical, 3 = Craft. 4 = Student, 5 = Homemaker, 6 = Retired, 7 = Self-employed |
| 58 | OCC\_Label | Unified Occupation, Labels |  |
| 59 | OWNRENT | Home ownership is missing |  |
| 60 | MARRYUN | Marital status unknown |  |
| 61 | MARRYYES | Married |  |
| 62 | MARRYNO | Not Married |  |
| 63 | MARRY | Marital status -unified | 0 = Unmarried, 1 = Married, 2 =Unknown |
| 64 | MARRY\_LABEL | Unified marital status, Labels |  |
| 65 | MAILORD | Buys via mail order |  |
| 66 | MAILRES | Responds to mail offers |  |
| 67 | MAILFLAG | Has chosen not to be solicited by mail |  |
| 68 | TRAVEL | Has traveled to non-US country |  |
| 69 | PCOWN | Owns a personal computer |  |
| 70 | CREDITCD | Possesses a credit card |  |
| 71 | RETCALLS | Number of calls previously made to retention team |  |
| 72 | RETACCPT | Number of previous retention offers accepted |  |
| 73 | NEWCELLY | Known to be a new cell phone user |  |
| 74 | NEWCELLN | Known not to be a new cell phone user |  |
| 75 | REFER | Number of referrals made by subscriber |  |
| 76 | INCMISS | Income data is missing |  |
| 77 | INCOME | Income (0=>missing) |  |
| 78 | MCYCLE | Owns a motorcycle |  |
| 79 | CREDITAD | Number of adjustments made to customer credit rating (up or down) |  |
| 80 | SETPRCM | Missing data on handset price |  |
| 81 | SETPRC | Handset price (0=>missing) |  |
| 82 | RETCALL | Customer has made made call to retention team | The retention team handles calls from customers considering whether to renew, reporting competitive offers, and so on. Retention team may make a retention offer to encourage customer to stay |
| 83 | CALIBRAT | Calibration sample = 1; Validation sample = 0; | Indicator of whether customer is in calibration or validation sample |
| 84 | CHURNDEP | Churn (=missing for validation sample) | Churn variable to use as dependent variable for logistic regression |

## Employee Attrition

The human resources department within GE has recently become aware that many high-potential employees have left the company to pursue other opportunities. This awareness was raised by many middle managers and supported by the recent increase in job postings. Given the need to remain competitive and the total cost and time required to train new employees, a need is present to identify how talent may be retained within GE. In the current environment the average cost of attrition for an individual is 80% of their annual salary.

The current data environment is an HR web-based desktop system that contains information about all employees, current and past, including their attrition status of YES—they have left—or NO—they have not left. This envornment includes metadata about each employee. It is maintained by HR staff via a web-based Java client server application. The data is stored in an Oracle database in transactional form. The IT department has built a data warehouse that is updated each evening with the current day data. The HR team uses this data warehouse data as the source of their reports and has the ability to have an ad hoc extract to select data into Excel for their unique research needs. The HR extracts are limited in rows as well as fields that have been corporately pre-approved for extraction.

GE has compiled a file for the purpose of this pilot project using the extract tool. The dataset provided in the **Assignment Guidelines and Rubrics** section of the course, Employee Attrition (CSV), includes information on employees which the GE HR team believes to be relevant to analyzing this problem.

The HR team would like to determine if this data can be used to identify attrition of employees that may leave. It is important to be able to understand attrition drivers for metadata like high performers, role types, and other pertinent groupings which come from the analysis.

The pilot will need to only show basis for this data to be able to describe and generally identify employees that may leave. The management team expects to make a GO or NO GO business decision based on the pilot recommendation. If there is a GO, then GE will allocate new project dollars to arrange for GE resources to develop a full-enterprise deployed predictive analytic model. Note that the results of this pilot will be used as a basis for that next project.

## Data Description/Documentation for Employee Attrition

|  |  |
| --- | --- |
| **Name** | **Description** |
| AGE | Numerical Value |
| ATTRITION | Employee leaving the company (0=no, 1=yes) |
| BUSINESS TRAVEL | (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely) |
| DAILY RATE | Numerical Value - Salary Level |
| DEPARTMENT | (1=HR, 2=R&D, 3=Sales) |
| DISTANCE FROM HOME | Numerical Value - THE DISTANCE FROM WORK TO HOME |
| EDUCATION | Numerical Value |
| EDUCATION FIELD | (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL) |
| EMPLOYEE COUNT | Numerical Value |
| EMPLOYEE NUMBER | Numerical Value - EMPLOYEE ID |
| ENVIROMENT SATISFACTION | Numerical Value - SATISFACTION WITH THE ENVIROMENT |
| GENDER | (1=FEMALE, 2=MALE) |
| HOURLY RATE | Numerical Value - HOURLY SALARY |
| JOB INVOLVEMENT | Numerical Value - JOB INVOLVEMENT |
| JOB LEVEL | Numerical Value - LEVEL OF JOB |
| JOB ROLE | (1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE) |
| JOB SATISFACTION | Numerical Value - SATISFACTION WITH THE JOB |
| MARITAL STATUS | (1=DIVORCED, 2=MARRIED, 3=SINGLE) |
| MONTHLY INCOME | Numerical Value - MONTHLY SALARY |
| MONTHY RATE | Numerical Value - MONTHY RATE |
| NUMCOMPANIES WORKED | Numerical Value - NO. OF COMPANIES WORKED AT |
| OVER 18 | (1=YES, 2=NO) |
| OVERTIME | (1=NO, 2=YES) |
| PERCENT SALARY HIKE | Numerical Value - PERCENTAGE INCREASE IN SALARY |
| PERFORMANCE RATING | Numerical Value - ERFORMANCE RATING |
| RELATIONS SATISFACTION | Numerical Value - RELATIONS SATISFACTION |
| STANDARD HOURS | Numerical Value - STANDARD HOURS |
| STOCK OPTIONS LEVEL | Numerical Value - STOCK OPTIONS |
| TOTAL WORKING YEARS | Numerical Value - TOTAL YEARS WORKED |
| TRAINING TIMES LAST YEAR | Numerical Value - HOURS SPENT TRAINING |
| WORK LIFE BALANCE | Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE |
| YEARS AT COMPANY | Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY |
| YEARS IN CURRENT ROLE | Numerical Value -YEARS IN CURRENT ROLE |
| YEARS SINCE LAST PROMOTION | Numerical Value - LAST PROMOTION |
| YEARS WITH CURRENT MANAGER | Numerical Value - YEARS SPENT WITH CURRENT MANAGER |