IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

INTRODUCTION

LECTURE 01

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

INTRODUCTION

- What is Business Analytics?
  - "Business analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states." - Gartner IT Glossary
  - Includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user
- Analytics can be classified as:
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics

INTRODUCTION

- Descriptive analytics involve gathering, organizing, tabulating, and depicting data and then describing the characteristics of what you are studying.
  - Also called reporting in managerial lingo
  - First phase of analytics
  - Though useful, it doesn't inform you about why the results happen or what can happen in future.

INTRODUCTION

- Predictive analytics use the past to predict the future.
  - Identify associations among different variables and predict the likelihood of a phenomenon reoccurring on the basis of those relationships
- Correlation vs. Causation
- Prescriptive analytics suggest a course of action.
  - Recommends decisions entailing mathematical and computational models
  - Final phase of analytics

INTRODUCTION

- Methods from statistics, forecasting, data mining, experimental design are used in Business Analytics
- What is Data Mining?
  - “Extracting useful information from large datasets” - Hand et al. 2001
  - “The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques.” - Gartner IT Glossary

INTRODUCTION

- Where is Data Mining Used?
  - Variety of domains:
    - Predicting the response of a drug or a medical treatment on the patient suffering from a serious disease
    - Predicting whether an intercepted communication is about a potential terror attack
    - Predicting whether a packet of network data can pose a cybersecurity threat

INTRODUCTION

- Where is Data Mining Used?
  - Common business questions:
    - Which customers are most likely to respond to the marketing or promotional offer?
    - Which customers are most likely to default on loan?
    - Which customers are most likely to subscribe to a magazine?

INTRODUCTION

- Data Mining Genesis
  - An interdisciplinary subfield of computer science
  - Originates from the fields of machine learning and statistics
  - Data mining as “statistics at scale and speed” – Pregibon 1999
  - Extension: “statistics at scale, speed, and simplicity” – Shmueli et al. 2010

INTRODUCTION

Classical Statistical Setting

- Data scarcity and Computational difficulty
- Same sample is used to compute an estimate and to check its reliability
- Logic of inference: confidence intervals and hypothesis tests (Inference is determining whether a pattern or result might have happened by chance)

Data Mining Paradigm

- Large datasets and fast computing powers
- Fitting a model with one sample and evaluating the performance using another sample
- Machine learning techniques, such as trees and neural networks are less structured and more computationally intensive in comparison to statistical techniques

INTRODUCTION

- Rapid Growth of Data
  - Millions of transactions on a daily basis
    - Organized retailers such as Shoppers Stop, Big Bazaar, and Pantaloons
    - E-commerce retailers such as Flipkart, Amazon, and Snapdeal
  - Growing economy and Internet growth
  - Decreasing cost and increasing availability of automatic data capture mechanisms, e.g., Bar codes, POS devices, click-stream data, GPS data
  - Operational databases to data warehouse and data marts
  - Constant declining cost of data storage and improving processing capabilities

INTRODUCTION

- Core of this course focuses on
  - Predictive Analytics consisting of tasks of
    - Prediction,
    - Classification,
    - Association rules

- In Data mining, typically several different methods are applied for a particular goal and the most useful is selected

INTRODUCTION

- Usefulness of a method
  - Goal of the analysis
  - Underlying assumptions of the method
  - Size of the dataset
  - Types of pattern in the dataset
- Dataset Example: Sedan Car owner
  - Goal: Income level and Household Area is used to classify whether a household owns a sedan car

INTRODUCTION

- Dataset format
  - Tabular or matrix format: variables in columns and observations in rows
  - Each row represents a household (unit of analysis) in SedanCar dataset
- R and RStudio
  - R is a programming language and software environment for statistical computing and graphics.
  - It is widely used by statisticians and data miners
  - RStudio is the most commonly used integrated development environment (IDE) for R.

INTRODUCTION

- Key Terms
  - Algorithm
    - A specific sequence of actions or set of rules to be followed to perform a task.
    - Algorithms are used to implement data mining techniques such as trees, neural networks etc.
  - Model
    - By model, we mean data mining model here
    - A data mining model is an application of a data mining technique on dataset

INTRODUCTION

- Key Terms
  - Variable
    - Operationalized way of representing a characteristic of an object, event, or phenomenon
    - A variable can take different values in different situations.
  - Input variable, Independent variable, Feature, Field, Attribute, or Predictor
    - Input variable is an input to the model

INTRODUCTION

- Key Terms
  - Output variable, Outcome variable , Dependent variable, Target variable, or Response
    - Output variable is an output of the model
  - Record, observation, case, row
    - Observation is the unit of analysis on which the variable measurements are taken such as a customer, a household, an organization, an industry etc.

INTRODUCTION

- In Data Mining and related domains, generally two types of variables are used:
  - Categorical
    - Nominal
    - Ordinal
  - Continuous
    - Interval
    - Ratio

INTRODUCTION

- Understanding the type of variables in a dataset is important
  - To identify an appropriate statistical or data mining technique
  - Proper interpretation of the data analysis results
- Data of these variable types are either quantitative or qualitative in nature
  - Quantitative data measure numeric values and are expressed in number
  - Qualitative data measure types and are expressed by a label, or a numeric code

INTRODUCTION

- Structure of these variable types increases from nominal to ratio in a hierarchical fashion

- Nominal
  - Values indicate distinct types, e.g., gender, nationality, religion, PIN code, employee ID
  - Only two operations = and ≠ are supported

INTRODUCTION

- Ordinal
  - Values indicate a natural order or sequence, e.g., academic grades, Likert scale, quality of a food item
  - Four additional operations $<$, $\leq$, $>$, $\geq$ are supported
- Interval
  - Difference between two values is also meaningful
  - Values may be in reference to a somewhat arbitrary zero point
  - Celsius temperature, Fahrenheit temperature, location variables: Distance from landmarks, geographical coordinates (latitude & longitude), calendar dates

INTRODUCTION

- Interval
  - Two additional operations +, - are supported
- Ratio
  - Ratio of two values is also meaningful. Values are in reference to an absolute zero point
  - Kelvin temperature, age, length, weight, height, income
  - Two additional operations ×, ÷ are supported

INTRODUCTION

- Conversion from one variable type to other
  - High structure variable type can be converted into low structure variable type
  - For example, a ratio variable ‘age’ can be converted into an ordinal variable ‘age group’

INTRODUCTION

- Course Roadmap
  - Module I: General Overview of Data Mining and its Components
  - Module II: Data Preparation and Exploration
  - Module III: Performance Metrics and Assessment
  - Module IV: Supervised Learning Methods
  - Module V: Unsupervised Learning Methods
  - Module VI: Time Series Forecasting
  - Module VII: Conclusion

INTRODUCTION

- Supplementary Lectures
  - Introduction to R
  - Basic Statistical Methods

Key References

- HBR Video (Business Analytics Defined by Thomas H. Davenport)
- Gartner IT Glossary
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

DATA MINING PROCESS

LECTURE 02

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DATA MINING PROCESS

- Phases in a typical Data Mining effort:
  1. Discovery

     Frame business problem

     Identify analytics component

     Formulate initial hypotheses

  2. Data Preparation

     Obtain dataset form internal and external sources

     Data consistency checks in terms of definitions of fields, units of measurement, time periods etc.,

     Sample

DATA MINING PROCESS

- Phases in a typical Data Mining effort:
  3. Data Exploration and Conditioning

     Missing data handling, Range reasonability, Outliers,

     Graphical or Visual Analysis

     Transformation, Creation of new variables, and Normalization

     Partitioning into Training, Validation, and Test datasets

DATA MINING PROCESS

- Phases in a typical Data Mining effort:

4. Model Planning

   Determine data mining task such as prediction, classification etc.

   Select appropriate data mining methods and techniques such as regression, neural networks, clustering etc.

5. Model Building

   Building different candidate models using selected techniques and their variants using training data

   Refine and select the final model using validation data

   Evaluate the final model on test data

DATA MINING PROCESS

- Phases in a typical Data Mining effort:
  6. Results Interpretation

     Model evaluation using key performance metrics
  7. Model Deployment

     Pilot project to integrate and run the model on operational systems
- Similar data mining methodologies developed by SAS and IBM Modeler (SPSS Clementine) are called SEMAA and CRISP-DM respectively

DATA MINING PROCESS

- Data mining techniques can be divided into Supervised Learning Methods and Unsupervised Learning Methods
- Supervised Learning
  - In supervised learning, algorithms are used to learn the function ‘f’ that can map input variables (X) into output variables (Y)

$$Y = f(X)$$

  - Idea is to approximate ‘f’ such that new data on input variables (X) can predict the output variables (Y) with minimum possible error (Ɛ)

DATA MINING PROCESS

- Supervised Learning problems can be grouped into prediction and classification problems
- Unsupervised Learning
  - In Unsupervised Learning, algorithms are used to learn the underlying structure or patterns hidden in the data
- Unsupervised Learning problems can be grouped into clustering and association rule learning problems

DATA MINING PROCESS

- Target Population
  - Subset of the population under study
  - Results are generalized to the target population
- Sample
  - Subset of the target population
- Simple Random Sampling
  - A sampling method wherein each observation has an equal chance of being selected

DATA MINING PROCESS

- Random Sampling
  - A sampling method wherein each observation does not necessarily have an equal chance of being selected
- Sampling with Replacement
  - Sample values are independent
- Sampling without Replacement
  - Sample values aren’t independent

DATA MINING PROCESS

- Sampling results in less no. of observations than the no. of total observations in the dataset
- Data Mining algorithms
  - Varying limitations on number of observations and variables
- Limitations due to computing power and storage capacity
- Limitations due to statistical software being used
- How many observations to build accurate models?

DATA MINING PROCESS

- Rare Event, e.g., low response rate in advertising by traditional mail or email
  - Oversampling of ‘success’ cases
  - Arises mainly in classification tasks
  - Costs of misclassification
    - Asymmetric costs due to more importance of ‘success’ class
  - Costs of failing to identify ‘success’ cases are generally more than costs of detailed review of all cases
  - Prediction of ‘success’ cases is likely to come at cost of misclassifying more ‘failure’ cases as ‘success’ cases than usual

DATA MINING PROCESS

- Dummy coding for categorical variables
  - Some statistical software cannot use categorical variables expressed in the label format
  - Dummy binary variables (having 0's and 1's: 0 indicating 'absence' and 1 indicating 'presence') for different classes of categorical variables are created
  - For example, if 'activity status' of individuals can be put into four mutually exclusive and jointly exhaustive classes as {student, unemployed, employed, retired}, only three dummy variables would be required

DATA MINING PROCESS

- Principle of Parsimony
  - A model or theory with less no. of assumptions and variables but with high explanatory power is generally desirable
- More no. of variables also increase the sample size requirements due to reliability of estimate
- Overfitting
  - A model built using a complex function that fits the data perfectly
  - Model ends up fitting the noise and explaining the chance variation

DATA MINING PROCESS

- Overfitting
  - More no. of iterations resulting in excessive learning of the data
  - More no. of variables in the model may lead to fitting spurious relationships
- Sample Size
  - Domain Knowledge
  - General rule of thumb: 10 × p observations, where p is the no. of predictors
  - For classification tasks: 6 × m × p observations, where m is the no. of classes in the outcome variable (Delmaster & Hancock, 2001)

DATA MINING PROCESS

- Outliers
  - A distant data point
  - Valid point or erroneous value?
  - Further review
    - Manual Inspection (Sorting, minimum and maximum values, clustering etc.)
    - Domain Knowledge
- Missing Values
  - Few records with missing values can be removed
  - Imputation

DATA MINING PROCESS

- Missing Values
  - Drop the variables having missing values
  - Replace with proxy variable
- Normalization
  - Standardization using z-score
  - Min-max normalization

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

INTRODUCTION TO R

LECTURE 03

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

INTRODUCTION to R

- Installation Steps for Windows PC or Laptop
  - Install R
    - Download Link: https://cran.r-project.org/bin/windows/base/
  - Install RStudio Desktop
    - Download Link: https://www.rstudio.com/products/rstudio/download/
  - Install JAVA if not already installed
    - Download Link: https://www.java.com/en/download/
- Open RStudio
  - Installing R packages

INTRODUCTION to R

- R Graphical User Interface (GUI)
  - Command-line interface (CLI)
  - Similar to BASH shell in LINUX or interactive version of scripting language Python
  - RStudio is a popular GUI for R and it has been used to write R scripts for this course

INTRODUCTION to R

- RStudio has four main window sections
  - Top-Left Section: To write and save R code (Script section)
  - Bottom-Left Section: To execute R code and output (Console section)
  - Top-Right Section: To manage datasets and variables (Data section)
  - Bottom-Right Section: To display plots and seek help on R functions (Plot and Help Section)

INTRODUCTION to R

- Dataset Import
  - In this course, datasets are either imported from Excel files or created in RStudio

- Open Rstudio
  - R Basics

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Basic Statistics Using R

LECTURE 04

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

Basic Statistics

- Descriptive Statistics
  - Open RStudio
- Hypothesis Testing
  - Formulate an assertion and test it using data
    - Comparing populations, e.g., comparing performance of students in exams for two different class sections
    - Testing the difference of the means from two data samples
  - A common technique to assess the difference or significance of the same

Basic Statistics

- Common assumption in Hypothesis testing
  - No difference between two samples
  - Referred as NULL Hypothesis H_0
  - Alternative Hypothesis (H_A): There is difference between two samples
- Example:
  - H_0: Students from class A and B had same performance in the examinations
  - H_A: Students from class A performed better than students from class B

Basic Statistics

- Hypothesis test leads to:
  - Either rejection of the null hypothesis in favor of the alternative
  - Or acceptance of the null hypothesis
- Examples:
  - H_0: New data mining model does not predict better than existing model
  - H_A: New data mining model predicts better than existing model

Basic Statistics

- Examples:
  - H_0: Regression coefficient is zero, i.e., variable has no impact on outcome
  - H_A: Regression coefficient is nonzero, i.e., variable has an impact on outcome
- A typical hypothesis test is comparing the means of two populations
- Normal Distribution
  - A common continuous probability distribution and useful due to Central limit theorem

Basic Statistics

- Difference of Means
  - Drawing inferences on two populations: P1 and P2
  - Compare means: μ_1 and μ_2
  - H_0: $\mu_1 = \mu_2$
  - H_A: $\mu_1 \neq \mu_2$
  - Basic approach: compare observed sample means: $\bar{x}_1$ and $\bar{x}_2$
- Student's t-test
  - Assumptions: Two population distributions (P1 and P2) have equal but unknown variances
  - Two samples of n_1 and n_2 observations drawn randomly and independently from P1 and P2, respectively

Basic Statistics

- Student's t-test
  - If P1 and P2 are normally distributed with same mean and variance
  - Then t-statistic follows a t-distribution with n_1+n_2-2 degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } {S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}$$

Basic Statistics

- Student's t-test
  - S_p is pooled standard deviation, S_1 and S_2 are sample standard deviation
  - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
  - Numerator of t is the difference of the sample means
    - Observed t value of 0 indicates the sample results are exactly equal to H_0
    - Observed t value being far enough from 0 and t-distribution indicating a low enough probability (<0.05) will lead to rejection of H_0
    - t-value falling in corresponding areas in the curve less than 5% of the time

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Basic Statistics Using R Part-2

LECTURE 05

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

Basic Statistics

- Student's t-test
  - If P1 and P2 are normally distributed with same mean and variance
  - Then t-statistic follows a t-distribution with n_1+n_2-2 degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } S_p{}^2 = \frac{(n_1-1)S_1{}^2+(n_2-1)S_2{}^2}{n_1+n_2-2}$$

Basic Statistics

- Student's t-test
  - S_p is pooled standard deviation, S_1 and S_2 are sample standard deviation
  - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
  - Numerator of t is the difference of the sample means
    - Observed t value of 0 indicates the sample results are exactly equal to H_0
    - Observed t value being far enough from 0 and t-distribution indicating a low enough probability (<0.05) will lead to rejection of H_0
    - t-value falling in corresponding areas in the curve less than 5% of the time

Basic Statistics

- Student's t-test
  - For a low probability, $\alpha = 0.05$, known as significance level of the test
  - t^* is determined such that $p(|t| \geq t^*) = \alpha$
  - H_0 is rejected if observed value of t is such that $|t| \geq t^*$
- Significance level of a statistical test is the probability of rejecting the null hypothesis
  - If null hypothesis is true and $\alpha = 0.05$, the observed magnitude of t would exceed t^* 5% of the time

Basic Statistics

- p-value is sum of p(t≤-|observed t-value|) and p(t≥|observed t-value|)
- Open Rstudio
- Welch's t-test
  - Used when assumption of equal population variance is not reasonable

$$t_w = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}}$$

Basic Statistics

- Welch's t-test
  - Assumption of random samples drawn from two normal populations with the same mean is still applicable
  - t-distribution
- Open RStudio
- Confidence Interval
  - Provide interval estimate of a population parameter using sample data
  - Indicates uncertainty associated with a point estimate
  - How close $\bar{x}$ is to μ

Basic Statistics

- Confidence Interval
  - A 95% confidence interval estimate for a population mean straddles the true unknown mean 95% of the time

$$\mu \in \bar{x} \pm \frac{2\sigma}{\sqrt{n}}$$

- Type I and Type II Errors

	H_0 is true	**H_0 is false**
H_0 accepted		Type II error
H_0 rejected	Type I error	

Basic Statistics

- Type I and Type II Errors
  - Significance level = type I error (Denoted by α)
    - Can be managed using appropriate significance level
  - Type II error (Denoted by β)
    - Can be managed using appropriate sample size
- Power of a test
  - Correctly rejecting H_0
  - $1-\beta$
  - Used to determine the sample size

Basic Statistics

- ANOVA
  - Used for more than two populations or groups instead of performing multiple t-tests
  - Generalization of hypothesis testing that is used for the difference of two group means
  - For n groups, n(n-1)/2 t-tests would be required
  - Multiple t-tests
    - Cognitively difficult
    - Increased probability of type I error

Basic Statistics

- ANOVA
  - H_0: All the population means are equal
  - H_A: At least one pair of the population means is not equal
  - Assumption: Each population is normally distributed with same variance
  - Test whether different population clusters are more tightly grouped or spread across all the populations

Basic Statistics

- ANOVA
  - Between-groups mean sum of squares (S_B^2)
    - An estimate of between-groups variance

$$S_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_0)^2$$

Where k=no. of groups, n_i is no. of observations in ith group, $\bar{x}_0$ is mean of all the groups, $\bar{x}_i$ is mean of ith group

  - Within-group mean sum of squares (S_W^2)
    - An estimate of within-group variance

Basic Statistics

- ANOVA
  - Within-group mean sum of squares (S_W^2)

$$S_W{}^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(x_{ij} - \bar{x}_i)^2$$

  - If $S_B^2 > S_W^2$, some of the population means are different
  - F-test statistic

$$F = \frac{S_B{}^2}{S_W{}^2}$$

- Open RStudio

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PARTITIONING PROCESS

LECTURE 6

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DATA MINING PROCESS

- Partitioning
  - Using same data for model building and model evaluation introduces bias
  - Selection of best model from several candidate models could be due to
    - Genuine superiority of the final model over other candidate models
    - Chance occurrence leading to better match between final model and data
  - Many data-driven techniques can end up producing the latter situation due to overfitting

DATA MINING PROCESS

- Partitioning
  - Partitioning of dataset into two or three parts can solve this problem
  - Typically three partitions- training, validation, and test sets are created following a predetermined proportions for each set and records are randomly assigned to different partitions
  - Sometimes records are assigned based on a relevant variable

DATA MINING PROCESS

- Partitioning
  - Training Partition
    - Usually largest
    - To build the candidate models
  - Validation Partition
    - To evaluate the candidate models
    - Or to fine-tune and improve the model
  - Test Partition
    - To evaluate the final model

DATA MINING PROCESS

- Types of Datasets
  - Cross-Sectional Data
    - Observations on variables related to many subjects (individuals, firms, industries, or countries)
    - Observed at same point of time (snapshot)
    - Unit of analysis is specified
    - Each observation represents a distinct subject
    - Main idea is to compare differences among the subjects

DATA MINING PROCESS

- Types of Datasets
  - Time Series Data
    - Observations on a variable related to one subject
    - Observed over a successive equally spaced points in time
    - Each observation represents a distinct time period
    - Main idea is to examine changes in the subject over time

DATA MINING PROCESS

- Types of Datasets
  - Panel Data or Longitudinal Data
    - Observations on variables related to same subjects over a successive equally spaced points in time
    - Main idea is to compare differences among the subjects and to examine changes in the subjects over time
    - Cross-sections with time order

DATA MINING PROCESS

- Types of Datasets
  - Pooled Cross-Sectional Data
    - Observations on variables related to subjects at different time periods
    - Main idea is to examine the impact on subjects due to environmental changes caused by certain events or policies
    - Independent cross-sections from different time periods

DATA MINING PROCESS

- Model Building
  - An example with Linear Regression
  - Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES

LECTURE 07

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- "A picture is worth a thousand words"
  - A popular proverb
- Primary Role in the Data Mining Process is during the Data Exploration, and Conditioning phase
  - Understanding data structure
  - Identifying gaps or erroneous values
  - Identifying outliers
  - Finding patterns

VISUALIZATION TECHNIQUES

- Primary Role in the Data Mining Process is during the Data Exploration, and Conditioning phase
  - Finding missing values
  - Identifying duplicate rows and columns
  - Variable selection, transformation and derivation
    - Appropriate bin sizes for converting continuous variable into categorical variable
    - Combining categories
    - Usefulness of variables and metrics

VISUALIZATION TECHNIQUES

- Data Exploration and Conditioning
  - Required preliminary step before formal analysis
  - Visual analysis
    - A free-form data exploration
    - Main idea is to support the data mining goal and subsequent formal analysis
    - Techniques range from basic plots to interactive visualizations
    - Features such filtering, zooming, color and multiple panels
  - Usage of Visualization Techniques depends on
    - Different data mining tasks such as classification, prediction, clustering etc.
    - Different data mining techniques such as CART, HAC etc.

VISUALIZATION TECHNIQUES

- Basic Charts
  - Display one or two variables at a time
  - Useful to understand the structure of the data, variable types, and missing values in the dataset
  - For Supervised learning methods, main focus is on outcome variable
    - Typically plotted on y-axis

VISUALIZATION TECHNIQUES

- Line Charts or Graphs
  - Used mainly to display time series data
  - Overall level and Changes over time
  - Open RStudio
- Bar Charts
  - For comparing groups using a single statistic
  - X-axis is used for categorical variable
  - Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES Part-2

LECTURE 08

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- Scatterplot
  - Useful for prediction tasks
    - Focus is on finding meaningful relationships between numerical variables
  - Useful for unsupervised learning tasks such as clustering
    - Focus is on finding information overlap
  - Both the axis are used for numerical variable
  - Open RStudio

VISUALIZATION TECHNIQUES

- Distribution Plots
  - Histogram and Boxplot
    - Distribution of a numerical variable
    - Directions for new variable derivations
    - Directions for binning of a numerical variable
  - Useful in supervised learning, specifically prediction tasks
    - Variable transformation in case of a skewed distribution
    - Selection of appropriate data mining method

VISUALIZATION TECHNIQUES

- Boxplots
  - Display entire distribution
  - Side-by-side boxplots for comparing groups
    - Importance of numerical predictors in classification tasks
  - Series of boxplots for changes in distributions over time
  - Open RStudio

VISUALIZATION TECHNIQUES

- Histograms
  - Display frequencies covering all the values
  - Vertical Bars are used
  - Open RStudio
- Heatmaps
  - Display numeric variables using graphics based on 2-D tables
    - Color schemes are used to indicate values
  - Useful to visualize correlation and missing values
    - Specially, in case of large no. of values

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- Open RStudio

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES Part-3

LECTURE 09

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- Histograms
  - Display frequencies covering all the values
  - Vertical Bars are used
  - Open RStudio
- Heatmaps
  - Display numeric variables using graphics based on 2-D tables
    - Color schemes are used to indicate values
  - Useful to visualize correlation and missing values
    - Specially, in case of large no. of values

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES Part-4

LECTURE 10

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Multiple panels
  - Color
  - Size and shape
  - Animation
  - Aggregation, rescaling, and Interactivity
  - Main idea is to help build visual perception to support the subsequent analysis
- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES Part-5

LECTURE 11

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

VISUALIZATION TECHNIQUES Part-6

LECTURE 12

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

VISUALIZATION TECHNIQUES

- Multidimensional Visualization
  - Trend Lines
  - In-plot labels
  - Scaling Up
  - Multivariate plots
- Specialized Visualization
  - Network graphs (Network data)
  - Treemaps (Hierarchical data)
  - Map Charts (Geographical data)

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

DIMENSION REDUCTION TECHNIQUES

LECTURE 13

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DIMENSION REDUCTION TECHNIQUES

- Large no. of variables
  - Subsets of variables might be highly correlated
  - Computational issues
  - Costs of data preparation, exploration, and conditioning
  - Dimensionality (Principle of Parsimony)
- Dimension Reduction is also called as factor selection or feature extraction is some domains

DIMENSION REDUCTION TECHNIQUES

- Dimension Reduction Techniques
  - Domain Knowledge
  - Data Exploration Techniques
  - Data Conversion Techniques
  - Automated reduction Techniques
  - Data Mining Techniques

DIMENSION REDUCTION TECHNIQUES

- Domain Knowledge
  - Identifying key variables for the data mining task
  - Removing redundant variables
  - Identifying erroneous variables
  - Measurement issues for variables

DIMENSION REDUCTION TECHNIQUES

- Data Exploration Techniques
  - Descriptive statistics
    - Summary statistics
    - Pivot tables
    - Correlation analysis
  - Visualization Techniques

- Open RStudio

DIMENSION REDUCTION TECHNIQUES

- Data Conversion Techniques
  - Combining categories
  - Converting a categorical variable into a numerical variable
- RStudio
- Automated reduction Techniques
  - Principal Component Analysis (PCA)

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

DIMENSION REDUCTION TECHNIQUES Part 2

LECTURE 14

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Used for reducing the no. of predictors
  - Used for quantitative variables
  - Highly correlated variable subsets
  - Main idea is to find a set of new variables that contains most of the information of original variables
  - Eliminating covariation and multicollinearity
  - Redistribution of variability
- Open RStudio

DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Data Mining Process
    - Apply PCA to the training partition
    - Predictors would now be principal score columns
    - Apply the principal weights obtained from training partition to the variables in the validation partition to obtain the scores
  - Relationship between predictors and output variable is ignored

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

DIMENSION REDUCTION TECHNIQUES Part-3

LECTURE 15

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Used for reducing the no. of predictors
  - Used for quantitative variables
  - Highly correlated variable subsets
  - Main idea is to find a set of new variables that contains most of the information of original variables
  - Eliminating covariation and multicollinearity
  - Redistribution of variability
- Open RStudio

DIMENSION REDUCTION TECHNIQUES

- Principal Component Analysis (PCA)
  - Data Mining Process
    - Apply PCA to the training partition
    - Predictors would now be principal score columns
    - Apply the principal weights obtained from training partition to the variables in the validation partition to obtain the scores
  - Relationship between predictors and output variable is ignored

DIMENSION REDUCTION TECHNIQUES

- Data Mining Techniques
  - Subset selection procedures using Regression models
    - Linear regression for prediction
    - Logistic regression for classification
    - Regression models can also be used for combining categories (using p-values)
  - Classification and Regression Tree (CART)
    - Classification tree for classification
    - Regression tree for prediction (Using tree diagram)

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS

LECTURE 16

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Need for performance metrics
  - Usefulness of a model
  - Comparison of candidate models
- Classification Performance
  - Probability of misclassification
  - Naïve Rule: most prevalent class
    - Serves as benchmark
  - Class Separation
- Open RStudio

PERFORMANCE METRICS

- Performance Metrics based on Naïve Rule
  - Multiple R^2
    - Distance between fit of model to data and fit of naïve rule to data
- Naïve rule equivalent for prediction
  - Sample mean
- Classification Matrix

  $n_{i,j}$: no. of class i cases classified as class j cases

Classification Matrix		
	Predicted Class	
Actual Class	1	0
1	$n_{1,1}$	$n_{1,0}$
0	$n_{0,1}$	$n_{0,0}$

PERFORMANCE METRICS

- Classification Performance
  - Validation partition classification matrix
  - Comparison of training partition classification matrix with validation partition classification matrix
    - Detect overfitting
- Performance Metrics based on classification matrix
  - Misclassification rate or error
  - Accuracy

PERFORMANCE METRICS

- Performance Metrics based on classification matrix

$$\text{err} = \frac{n_{0,1} + n_{1,0}}{n}$$

$$\text{accuracy} = 1 - \text{err} = \frac{n_{0,0} + n_{1,1}}{n}$$

- Open RStudio
- Cutoff probability value
  - Accuracy for all the classes is important
    - A case is assigned to the class with the highest probability as estimated by the model

PERFORMANCE METRICS

- Cutoff probability value
  - Accuracy for a particular class of interest is important
    - A case is assigned to the class of interest if probability for the class is above cutoff value
  - Default cutoff value for a two class model is 0.5 (principally similar to naïve rule)
- Open Excel
  - One-variable table

PERFORMANCE METRICS

- Why change cutoff probability value from 0.5?
  - Class of interest
  - Asymmetric misclassification cost

- When to incorporate change in cutoff value?
  - After final model selection
  - Before model derivation

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-2

LECTURE 17

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Performance Metrics in presence of a class of interest

$$\text{sensitivity} = \frac{n_{11}'}{n_{10}'+n_{11}'} = \text{true positive fraction}$$

$$\text{specificity} = \frac{n_{00}'}{n_{00}'+n_{01}'} = \text{true negative fraction}$$

- ROC (receiver operating characteristic) curve
  - Used to plot {sensitivity, 1-specificity} points as the cutoff value increases
  - Top left corner points reflect wanted performance

PERFORMANCE METRICS

- Open Excel and RStudio
- Rank Ordering of records for class of interest
  - Based on estimated probabilities of class membership

- Lift curve is used to display the effectiveness of the model in rank ordering of cases
  - Constructed using validation partition scores

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-3

LECTURE 18

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Performance Metrics in presence of a class of interest

$$\text{sensitivity} = \frac{n_{11}'}{n_{10}'+n_{11}'} = \text{true positive fraction}$$

$$\text{specificity} = \frac{n_{00}'}{n_{00}'+n_{01}'} = \text{true negative fraction}$$

- ROC (receiver operating characteristic) curve
  - Used to plot {sensitivity, 1-specificity} points as the cutoff value increases
  - Top left corner points reflect wanted performance

PERFORMANCE METRICS

- Open Excel and RStudio
- Rank Ordering of records for class of interest
  - Based on estimated probabilities of class membership

- Lift curve is used to display the effectiveness of the model in rank ordering of cases
  - Constructed using validation partition scores

PERFORMANCE METRICS

- Cumulative lift curve or gains chart
  - Used to plot cumulative no. of cases on x-axis and cumulative no. of true positive cases on y-axis
  - Plot displays the lift value of the model for a given no. of cases w.r.t the random selection (probability value of class membership determines the reference line)
- Open Excel and RStudio
- Decile Chart
  - Alternative plot to convey the same information as gains chart

PERFORMANCE METRICS

- Open RStudio
- Asymmetric Misclassification Costs
  - When misclassification error for a class of interest is more costly than for the other class
  - Example, misclassifying a customer as false positive who is actually likely to respond to the promotional offering
    - Opportunity cost of foregone sale vs. costs of making an offer (profit of ₹20 for a ₹ 100 item vs. ₹1 scenario)
  - Misclassification rate is not appropriate metric in this case

PERFORMANCE METRICS

- Asymmetric Misclassification Costs
  - Other considerations
    - Costs of analyzing data
    - Actual net value impact per record
    - New Goal : minimization of costs or maximization of profits
- Open Excel
- How to improve actual classifications by incorporating asymmetric misclassification costs?
    - Change the rules of classification e.g. cutoff value

PERFORMANCE METRICS

- Performance Metrics based on asymmetric misclassification costs

$$\text{average misclassification cost} = \frac{c_0 n_{0,1} + c_1 n_{1,0}}{n}$$

  - Measures average cost of misclassification per observation
  - Where c_i is cost of misclassifying a class i observation

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-4

LECTURE 19

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Open RStudio
- Asymmetric Misclassification Costs
  - When misclassification error for a class of interest is more costly than for the other class
  - Example, misclassifying a customer as false positive who is actually likely to respond to the promotional offering
    - Opportunity cost of foregone sale vs. costs of making an offer (profit of ₹20 for a ₹ 100 item vs. ₹1 scenario)
  - Misclassification rate is not appropriate metric in this case

PERFORMANCE METRICS

- Asymmetric Misclassification Costs
  - Other considerations
    - Costs of analyzing data
    - Actual net value impact per record
    - New Goal : minimization of costs or maximization of profits
- Open Excel
- How to improve actual classifications by incorporating asymmetric misclassification costs?
    - Change the rules of classification e.g. cutoff value

PERFORMANCE METRICS

- Performance Metrics based on asymmetric misclassification costs

$$\text{average misclassification cost} = \frac{c_0 n_{0,1} + c_1 n_{1,0}}{n}$$

  - Measures average cost of misclassification per observation
  - Where c_i is cost of misclassifying a class i observation

PERFORMANCE METRICS

- Ratio of costs (c_0/c_1)
- Future misclassification costs
  - Prior Probabilities (p_0/p_1)
  - (p_0/p_1)* (c_0/c_1)
- Lift curve incorporating costs
- Open RStudio
- Lift vs.
  - No. of records or cutoff value?

PERFORMANCE METRICS

- Asymmetric misclassification costs for m classes (m>2)
  - Classification matrix will be ‘m×m’
  - m prior probabilities
  - m(m-1) misclassification costs
  - Matrix for misclassification costs becomes complicated
  - Lift chart not usable for multiclass scenario

PERFORMANCE METRICS

- Oversampling of rare class members
  - Simple random sampling vs. stratified sampling
- Oversampling approach
  1. Sample more rare class observations (equivalent of oversampling without replacement)
     - Lack of adequate no. of rare class observations
     - Ratio of costs is difficult to determine
  2. Replicate existing rare class observations (equivalent of oversampling with replacement)

PERFORMANCE METRICS

- Typical solution adopted by analysts
  - Sample equal no. of members from both the classes
- Oversampling adjustment for performance evaluation
  - Score
    1. Validation partition without oversampling
    2. Oversampled validation partition and then remove the oversampling effects by adjusting weights

PERFORMANCE METRICS

- Typical steps in rare class scenario
  1. Build the candidate models on training partition with 50% class 1 observations and 50% class 0 observations
  2. Validate the models with the validation partition drawn using simple random sample taken from original dataset
- Detailed steps
  1. Separate the class 1 and class 0 observations into two strata (distinct sets)
  2. Half the records from class 1 stratum are randomly selected into training partition

PERFORMANCE METRICS

- Detailed steps
  3. Remaining class 1 records are reserved for validation partition
  4. Randomly select class 0 records for training partition equal to no. of class 1 records in step 2
  5. Randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition
  6. For test partition, a random sample can be taken from validation partition

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PERFORMANCE METRICS Part-5

LECTURE 20

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Oversampling of rare class members
  - Simple random sampling vs. stratified sampling
- Oversampling approach
  1. Sample more rare class observations (equivalent of oversampling without replacement)
     - Lack of adequate no. of rare class observations
     - Ratio of costs is difficult to determine
  2. Replicate existing rare class observations (equivalent of oversampling with replacement)

PERFORMANCE METRICS

- Typical solution adopted by analysts
  - Sample equal no. of members from both the classes
- Oversampling adjustment for performance evaluation
  - Score
    1. Validation partition without oversampling
    2. Oversampled validation partition and then remove the oversampling effects by adjusting weights

PERFORMANCE METRICS

- Typical steps in rare class scenario
  1. Build the candidate models on training partition with 50% class 1 observations and 50% class 0 observations
  2. Validate the models with the validation partition drawn using simple random sample taken from original dataset
- Detailed steps
  1. Separate the class 1 and class 0 observations into two strata (distinct sets)
  2. Half the records from class 1 stratum are randomly selected into training partition

PERFORMANCE METRICS

- Detailed steps
  3. Remaining class 1 records are reserved for validation partition
  4. Randomly select class 0 records for training partition equal to no. of class 1 records in step 2
  5. Randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition
  6. For test partition, a random sample can be taken from validation partition

PERFORMANCE METRICS

- When ‘Validation partition without oversampling’ is not useful
  - Due to very few class 1 records
  - Second approach of

  ‘Using oversampled validation partition for evaluation as well and adjusting the weights to get rid of oversampling effects’

  is taken
    - Adjustment of validation partition classification matrix and lift curve is performed to get reliable accuracy measures

PERFORMANCE METRICS

- Lift Curve on oversampled validation partition
  - Multiply the net value of a record with proportion of class 1 records in original data
- In a two-class scenario, records which are difficult to classify by the model, can be labeled with a third class option
  - ‘cannot say’
  - Expert judgment can be used for such cases

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

PREDICTION PERFORMANCE

LECTURE 21

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

PERFORMANCE METRICS

- Prediction Performance
  - Continuous outcome variable
  - Predictive accuracy vs. goodness of fit
  - E.g., goodness of fit measures in regression modeling
    - R^2 and std. err estimate
    - Residual analysis
  - Prediction Error on validation partition
  - Benchmark criterion: average

PERFORMANCE METRICS

- Prediction Error
  - For a record i, prediction error = actual value - predicted value
  - $e_i = y_i - \hat{y}_i$
- Predictive Accuracy Measures
  - Average Error

$$\frac{1}{n}\sum_{i=1}^{n} e_i$$

    - On average, indicates over or under prediction

PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Mean Absolute Error (MAE) or Mean Absolute Deviation (MAD)

$$\frac{1}{n}\sum_{i=1}^{n}|e_i|$$

    - On average, magnitude of error
  - Mean Absolute Percentage Error (MAPE)

$$100\% \times \frac{1}{n}\sum_{i=1}^{n}|e_i/yi|$$

    - On average, percentage deviation from actual values

PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

  - Similar to std. err estimate computed on validation partition
  - Measured in same unit as the outcome variable

PERFORMANCE METRICS

- Predictive Accuracy Measures
  - Total Sum of Squared Errors (Total SSE or SSE)

$$\sum_{i=1}^{n} e_i^2$$

- These Predictive Accuracy Measures are used to
  - Compare the candidate models
  - Degree of prediction accuracy
  - Outlier issues

PERFORMANCE METRICS

- Outlier influence in accuracy measures
  - By comparing median based measures and mean based measures
  - Histogram or boxplot of residuals
- Model with high predictive accuracy may or may not be same as
  - model with best fit of data
- Evaluation using visualization techniques
  - Lift curve
    - Relevant when records with highest predicted values are sought

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION

LECTURES 22

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Most popular model
- Idea is to fit a linear relationship between
  - A quantitative outcome variable (Y) and
  - A set of p predictors $\{X_1, X_2, X_3, \ldots, X_p\}$
- Assumption: relationship as expressed in the following model equation holds true for the target population

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where $\beta_0, \ldots, \beta_p$ are coefficients and ε is the noise or unexplained part

MULTIPLE LINEAR REGRESSION

- Objectives:
  - Understanding the relationships between outcome variable and predictors
    - Followed in statistical approach
  - Predicting values of outcome variable for new records
    - Followed in data mining approach
- Applications in data mining
  - Predicting credit card spending, life of an equipment, sales etc.

MULTIPLE LINEAR REGRESSION

- Model Building and Results Interpretation phases differ depending on the objective:
  - Explanatory (predicting the impact of promotional offer on sales)
  - Predictive (predicting sales)
- Selection of suitable data mining techniques depends on the goal itself

MULTIPLE LINEAR REGRESSION

Explanatory Modeling

- Fits the data closely
- Full sample is used to estimate best-fit model
- Performance metrics measure how close model fits the data

Predictive Modeling

- Predicts new records accurately
- Sample is partitioned into training, validation, and test sets and training partition is used to estimate the model
- Performance metrics measure how well model predicts new observations

MULTIPLE LINEAR REGRESSION

Explanatory Modeling

- Model might not have best predictive accuracy
- Statistical techniques with assumed or hypothesized relationships and scarce data (primary data)

Predictive Modeling

- Model might not be best-fit of data
- Machine learning techniques with no assumed structure and large datasets (secondary data)

MULTIPLE LINEAR REGRESSION

- Estimates for target population
  - Coefficients: β_0, …, β_p and
  - σ, std. deviation of noise (ε)
  - Cannot be measured directly due to unavailability of data on entire population
- Estimation technique:
  - Ordinary least squares (OLS)
    - Computes the sample estimates which minimize the sum of squared deviations between actual values and predicted values

MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

  - Unbiased predictions (on average, closer to actual values)
  - Smallest average squared error

  Given following assumptions hold true

  - Noise follows a normal distribution
  - Linear relationship holds true
  - Observations are independent
  - Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors

MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION Part-2

LECTURES 23

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

  - Unbiased predictions (on average, closer to actual values)
  - Smallest average squared error

  Given following assumptions hold true

  - Noise follows a normal distribution
  - Linear relationship holds true
  - Observations are independent
  - Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors

MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION Part-3

LECTURES 24

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

  - Unbiased predictions (on average, closer to actual values)
  - Smallest average squared error

  Given following assumptions hold true

  - Noise follows a normal distribution
  - Linear relationship holds true
  - Observations are independent
  - Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors

MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio

MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Availability of large no. of variables for selecting a set of predictors
  - Main idea is to select most useful set of predictors for a given outcome variable of interest
  - Selecting all the variables in the model is not recommended
    - Data collection issues in future
    - Measurement accuracy issues for some variables
    - Missing values
    - Parsimony

MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Selecting all the variables in the model is not recommended
    - Multicollinearity: two or more predictors sharing the same linear relationship with the outcome variable
    - Sample size issues: Rule of thumb

$$n > 5*(p+2)$$

Where n=no. of observations

And p=no. of predictors

    - Variance of predictions might increase due to inclusion of predictors which are uncorrelated with the outcome variable
    - Average error of predictions might increase due to exclusion of predictors which are correlated with the outcome variable

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION Part-4

LECTURES 25

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Availability of large no. of variables for selecting a set of predictors
  - Main idea is to select most useful set of predictors for a given outcome variable of interest
  - Selecting all the variables in the model is not recommended
    - Data collection issues in future
    - Measurement accuracy issues for some variables
    - Missing values
    - Parsimony

MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Selecting all the variables in the model is not recommended
    - Multicollinearity: two or more predictors sharing the same linear relationship with the outcome variable
    - Sample size issues: Rule of thumb

$$n > 5*(p+2)$$

Where n=no. of observations

And p=no. of predictors

    - Variance of predictions might increase due to inclusion of predictors which are uncorrelated with the outcome variable
    - Average error of predictions might increase due to exclusion of predictors which are correlated with the outcome variable

MULTIPLE LINEAR REGRESSION

- Bias-variance trade-off
  - too few vs. too many predictors
    - Few predictors -> higher bias -> lower variance
  - Drop variables with ‘coefficient < std. dev. of noise’ and with moderate or high correlation with other variables
    - Lower variance
- Steps to reduce the no. of predictors
  - Domain knowledge
  - Practical reasons

MULTIPLE LINEAR REGRESSION

- Steps to reduce the no. of predictors
  - Summary statistics and graphs
  - Statistical methods using computational power
    - Exhaustive search: all possible combinations
    - Partial-iterative search: algorithm based
- Exhaustive Search
  - Large no. of subsets
  - Criteria to compare models
    - Adjusted R^2

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…