Business School
UNIVERSITY OF COLORADO DENVER

Information Systems Program

# Module 6
# SQL for Data Mining Input

## Lesson 1: Motivation and Background

# Lesson Objectives

- Explain inputs and outputs of data mining for association rules and classification
- Discuss input requirements for association rules and classification
- Identify differences between data lakes and data warehouses for data mining input

# Motivation

- Prevalence of parent-child relationships for data mining algorithms
- Input formats sometimes difficult to generate in SQL
- Increasing interaction between data warehouse professionals and data scientists
- Improved SQL skills

# Association Rules

- Set of baskets containing items
- Occurrence rules
  - IF Item1, Item2, … ItemN-1 THEN ItemN
  - Evaluation measures to select best rules

| *BasketId* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

**Rules Discovered:**

**{Milk} --> {Coke}**
**{Diaper, Milk} --> {Beer}**

# Applications of Association Rule Mining

Market basket analysis using items purchased together

Medical data analysis using disease occurrences and symptoms, locations and frequent diseases, and treatments and complications

Insurance coverage with policies and covered items and types of policies purchased together

# Classification

| Training Data | | | | |
|---|---|---|---|---|
| *CustId* | *Age* | *Income* | *LoanAmt* | *Default* |
| 1 | 59 | $66,150 | $8,100 | False |
| 2 | 18 | 70,000 | $8,775 | True |
| 3 | 39 | $25,500 | $1,400 | False |

Classification Algorithm

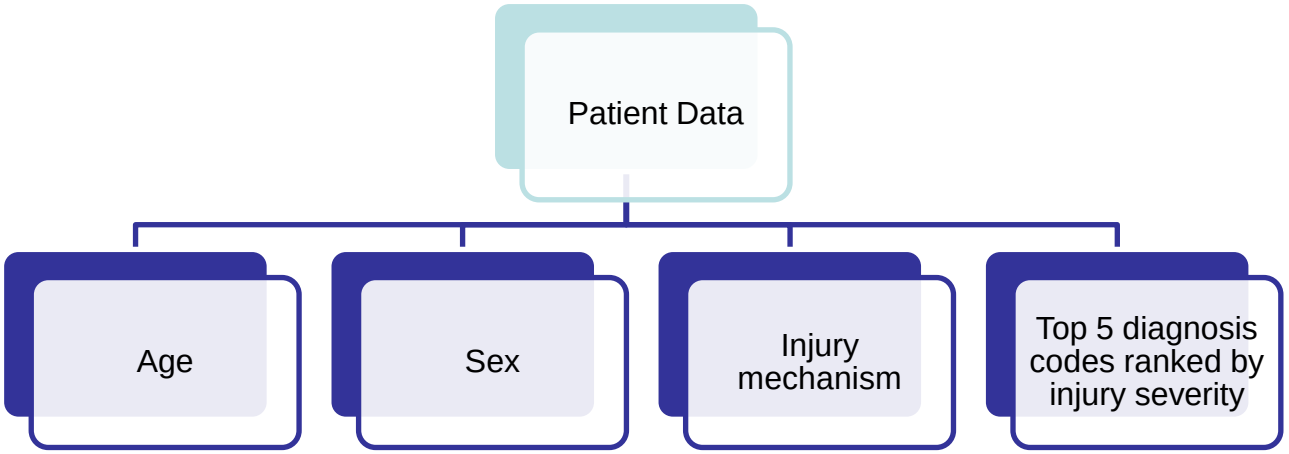
| Unseen Cases | | | |
|---|---|---|---|
| *CustId* | *Age* | *Income* | *LoanAmt* |
| 1100 | 61 | $68,200 | $10,100 |
| 1101 | 20 | $75,500 | $9,855 |
| 1102 | 35 | $25,500 | $2,500 |

Classification Model

| Predictions | | | | |
|---|---|---|---|---|
| *CustId* | *Age* | *Income* | *LoanAmt* | ***Prediction*** |
| 1100 | 61 | $68,200 | $10,100 | False |
| 1101 | 20 | $75,500 | $9,855 | True |
| 1102 | 35 | $25,500 | $2,500 | False |

# Retrospective Trauma Mortality Prediction

- Post evaluation of trauma center performance
- Predict death/survival for trauma patients
- Many studies over several decades



# Input Requirements

- Focus on SQL statements
- Data preparation and data reduction not SQL concern so not covered
- Difficult to generate input formats using SQL
- Flatten parent-child relationships into a single table

# Input Sources

## Data Lake

- Parent-child relationships
- Basket identifiers
- Derived directly from operational database

## Data Warehouse

- Fact table in relationships with dimension tables
- Sometimes lack basket identifiers
- Many transformations on data sources

# Summary

- Collaboration between data warehouse professionals and data scientists
- Prominence of data mining for association rules and classification with limited event history
- Work with data lakes or data warehouses
- Specialized but important skills extending beyond this course