Chapter 4: Intro to predictive Modeling

Difference between predictive descriptive (BI) and predictive analytics?

Descriptive – details about specific people  
Predictive – Low churn risk vs high churn risk

Supervised Learning – Learn from my example:  
In supervised systems, the data presented to a machine learning algorithm is fully labeled.

Unsupervised Learning – See what you can find:  
In unsupervised learning, systems are not provided any training examples at all and conduct clustering

Columns in a table – “Model Attributes” – Target Variables, what we are trying to predict.

Variable Types –

Numeric: Anything that has some order

(Numbers, Dates), Dimension of 1

Categorical: Stuff that does not have an order

Binary, text, Dimension = Number of possible values

Data Mining vs Use of the Model

Supervised modeling: Training the Model to predict results, then using the model to predict values.

Sample: Survivorship Issues

Lending Club want to have a model to take over the screening process that selects applications and deny those that are likely to default  
Provided: Past loans and outcomes

Bad:  
 - Use the data they currently have (Screened applicants) to predict default

Sample: Different Sources

Things go really bad if the positive and negative are treated differently  
Looking for drivers of diabetes: how do assemble the training data?

Bad:

* Go to a specialized hospital and get records from people treated for diabetes
* -Go somewhere else (Gym) to get records from healthy people

\*Training samples should be as similar as possible to the USE data

**Dimensionality of the data?**The sum of the number of numeric features and the number of the values of categorical features

**Classification**

Classification is a supervised learning technique which maps data into predefined classes or groups  
Training set contains a set of records, where one of the records indicates class.  
Modeling objective is to assign a class variable to all of the records, using attributes of other variables to predict a class  
Data is divided into test/train, where train is used to build the model and test is used to validate the accuracy of the classification  
Typical techniques: Decision Trees, Neural Networks

Example:

Gender, Age, Lipstick  
IE: Customers > Female / male? > 15 years+ / less than 15 years

**Classification: Creating Model**

Training Data -> Classification Algoriithms -> Trained classifier -> Purchased lipstick if female and age >= 15

**Classification: Applying Rules**if Gender = Female and Age >=15, then purchase lipstick = yes

**Selecting Informative Attributes**Objective: Based on customer attributes, patitirion the customers into subgroups that are less impure – with respect to the class (ie, such that in each group as many instances as possible belong to the same class)

**Multivariate Supervised Classification**  
If we select the single variable that gives the most information gain, we create a very simple segmentation.  
If we select multiple attributes each giving some information gain, how do we put them together?

**Tree Structure Models**Root node (Thickest Branch)  
 Interior Nodes  
Leaf Nodes

**Basic Rules for Tree Structured Models**

-No parents share descendants  
-No cycles  
- Branches always point downwards  
- Every example always ends up at a leaf node with some specific class determination.  
 - Probability estimation trees, regression trees

**How do we create a classification tree from data** - Divide and conquer  
 - take each data subset and recursively apply attribute selection to find the best attribute to partition it

**When do we stop?** The Nodes are pure, there are no more variables or even earlier to prevent over fitting.

**Why Trees / Tree Structured Models?**Decision trees (DTs) or classification trees are one of the most popular data mining tools, along with linear and logistic regression  
Easy to understand  
Easy to implement  
Easy to use  
Computationally cheap  
Almost all data mining packages include DTs  
They have advantages for model comprehensibility, which is important for:  
 Model Evaluation  
 Communication to non-DM-Savvy Stakeholders