PREDICTIVE ANALYTICS
(IBM SPSS Modeler)

2020-21
B. Tech 3rd year

Data Analysis
- Inspect, cleanse, transform & model data to discover useful info, informing conclusions & support decision-making.
- Analyze of raw data to make conclusions

Steps
- Determine data requirements/how data is grouped
- Collect data
- Organise data
- Cleaning data
  ↳ Analysis

Types

① Descriptive
- What happened?
- Summarise large datasets to describe outcomes.
- Help track success/failure
- Provide essential insights into past performance.

② Diagnostic
- Why things happened?
- Takes findings from descriptive analysis & dig deeper to find the cause
- Performance indicator (Identify anomalies in data, data resembling anomalies collect, statistical techniques to find relationships & trends to explain anomalies).

③ Predictive
- What's likely to happen in the future?
- Use historical data to identify trends & determine if they're likely to recur.
- Provide valuable insight as to what might happen in future using various statistical & ML techniques.

④ Prescriptive
- What's the best course of action
- Using insights from predictive analysis, data-driven decisions can be made.
- Allows to make informed decisions in face of uncertainty.
- Rely on ML strategies to find patterns in large datasets & likelihood of various outcomes estimated

Importance
- Help businesses optimize their performance
- Make informed decisions
- Identify efficient ways to complete tasks
- Analyze customer trends & satisfaction

Applications (TIMSHELMF)
- Transport & Delivery
- Logistics & Delivery
- Internet
- Manufacturing
- Security
- Education
- Healthcare
- Military
- finance

Data Mining
The process of extracting information to identify patterns, trends from raw data to convert it into useful information that would allow a business to take data-driven decisions from huge datasets.

Methodology
- CRISP-DM — Cross Industry Standard Process for Data Mining
- KDD — Knowledge Discovery in Databases
- SEMMA — Sample, Explore, Modify, Model, Access

Advantages
- Business Management → New opportunities
- Marketing Strategies
- Brand Strengthening
- Customer Identification
- Revenue Growth

Disadvantages
- Privacy issues
- Security issues
- Misuse of info/inaccurate info
- Cost

Applications
- financial data analysis
- Retail industry
- Telecom industry
- Biological Data Analysis
- Intrusion Detection

Skills required (DBUTECT)

- Understand business (ask right questions, evaluation of alternative sol^n)
- Database knowledge
- Knowledge of data-mining techniques
- Project management & experience
- Communication & Presentation skills
- Relevant technical background

Successes

- Result assessment w.r.t business success, not statistical criteria
- factors ⇒ customer satisfaction, reduced savings, profits, ROI
- Monitor model after deployment
- Cost of errors

failures

- Bad data
- Deployment can be hard
- Organizational resistance in deploying a sol^n
- Starting with wrong questions
- Lack of diverse expertise

CRISP-DM

- Open standard for clear model of analysis
- Provides roadmap to mine & analyze data & ↑ possibility of professional collaboration.
- Helps client understand what standards to expect.
- Phases are sequential & process is iterative.
- Stages influence each other in non-linear manner.

Stages (BDPMED)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Palettes (fSRfGMOEI)

- favorites
- Sources
- Record Ops
- field Ops
- Graphs
- Modeling
- Output
- Export

IBM SPSS Statistics

Pros

- flexible
- long-term strategy
- functional templates
- Cost-effective
- Implementation in any DS domain

Cons

- Heavy documenting
- Not a project mngmnt approach.
- Not modern.

Node

- Operations on data
- Linked together to form a stream.

Stream

- Flow of data from importing data, through a no. of manipulations, to running an analysis

SPSS Interface (MTCPMP)

- Main Menu
- Toolbar
- Stream Canvas
- Palettes
- Panes (Manager pane (Stream, Output, Models)
  Project pane (CRISP-DM, Classes))

Super Node

- Condense a no. of nodes into a single node, to make stream clean & manageable

Generate Node

- Multi-purpose node, having less functionality
- Same interface as select node

Select Node

- Create a combined compound condition joined by "AND" or "OR"
- Specific functionality & more operations

Sort Node

- Sort records in ascending or descending order based on values of one or more fields
- Upper field is given priority
- Only rows are rearranged & data is preserved for each row
- Value of actual data is not changed but condition & priority sorting is done for each field.

Sample Node

- Select subset of records for analysis or to discard
- Improve performance by estimating models on subsets
- Select group of related records for analysis
- Identify units for random inspection
- Stratified, clustered, structured.

Derive Node

- Modify data values / derive new fields from existing data
- Create fields of type formula, flag, nominal, state, count, Conditional
- Column name similar to name of derived node

Palettes & Their Nodes

① Sources

- Bring data into SPSS modeler

i). Var. file

- Read data from free-field text files (files whose records contain a constant no. of fields but varied no. of characters), also called delimited text files.
- Useful for files with fixed length header text & certain types of annotations.
- Records are read one at a time & passed through stream until entire file is read.
- field delimiters ⇒ space, comma, tab, newline, others.
- Also provide functionalities of filter & type nodes.

ii). Statistics file

- Reads data from .sav/.zsav file formats used by IBM SPSS Statistics, as well as cache file saved in IBM SPSS Modeler.
- file can be password protected & uses 2 method to handle variable names & labels (Read names & labels; Read labels as names)
- Also provides functionalities of filter & type nodes.

iii). Excel

- Enables to import data from Microsoft Excel in the .xlsx file format.
- We can choose the worksheet style (by name or index) and the range on worksheet.
- Provide functionalities of filter & type nodes.

② Record Ops

- Make changes at record level of the data
- Imp. during data understanding & preparation phases of data mining.
- Tailor the data acc. to specific business needs.

i). Select

- Select or discard a subset of records from data stream based on a specific condition
- Mode specifies whether records meeting condition will be included or excluded from data stream.
- Condition displays the selection criteria used to test each record, specified using a CLEM expression.
- Condition can be entered as an expression in window or use the Expression Builder by clicking calculator symbol.

ii). Sample

- Select a subset of records for analysis, or to specify a proportion of records to discard.
- A variety of sample types are supported viz., stratified, clustered & non-random (structured) samples.
- Used to
  - improve performance by estimating models on a data subset
  - select groups of related records for analysis
  - identify units/cases for random inspect for quality assurance
- Methods (first n, 1-in-n, Random %)

iii). Sort

- Sort records in ascending/descending order based on values of one or more fields.
- All fields selected to use as sort keys are displayed in table (key fields prefered to be numeric)
- Sort node is not applied if there's a Distinct node down the model flow.
- Upper key field is prioritized in sorting operations.
- Default sort order can be set as required.

iv). Distinct

- find/remove duplicate records in data and create a single, composite record from a group of duplicate records.
- A set of key fields must be defined to determine when 2 records are considered to be duplicates.
- If all fields are not picked as "key fields", then 2 duplicate records may not be truly identical.
- Sort order can also be applied within each group of duplicate records → gives control over which record is treated as the first one.

v). Aggregate 

- Aggregation is a data preparation task used to reduce the size of dataset
- Replace a sequence of input records with summary, aggregated output records.

vi). Merge 

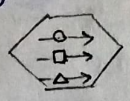
- Takes multiple input records & create a single output record containing all or some of the input fields
- Useful operation when merging of data from different sources is required.
- Merge methods (Order, keys, condition, etc.)
- Can also filter out columns as required

vii). Append 

- Concatenate set of records.
- It reads & pass downstream all records from one source until it is exhausted. Then, the records are read from next source using the same fields as input.
- for any incomplete values, system null string ($null$) is used.
- Used for combining datasets with similar structures but different data.

(3) field Ops

- Used to select, clean or construct data in prep for analysis

i). Type 

- Specify field properties (values, labels, etc.)
- Specify measurement level (type of field), its role in dataset (input, target), value mode, values check, etc.
- Define missing values

* Measurement levels

a. Default - storage values & type are unknown
b. Continuous - Describe numeric values
c. Categorical - String values when exact no of distinct values is known.
d. flag - Data with only 2 distinct values
e. Nominal - Data with multiple distinct values, each treated as a member of a set
f. Ordinal - Data with multiple distinct values, having an inherent order
g. Typeless - fields with a single value
Nominal data where set members > 250

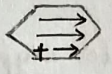
Instantiation

Process of reading/specifying info, such as storage type or values for a data field.

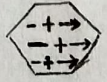
A user-directed process, you tell the software to read values by running data through a Type node.

ii). filter 

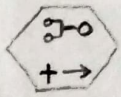
- Rename/exclude fields at any point.
- filter tab in available in various nodes to define or edit multiple response sets.
- Map fields from one input/import node to another.

iii). Derive 

- Modify data values & derive new fields from existing data.
- Create fields of type formula, flag, nominal, state, count, & conditional.
- You can also name the derived field as per your choice.

iv). filler 

- Replace field values & change storage
- Values can be replaced based on a specified CLEM condition (@BLANK(FIELD)) or all blanks/null values with a specific values.
- Often used in conjunction with Type Node to replace missing values.

v). Reclassify 

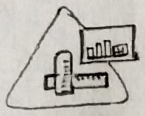
- Enables transformation from one set of categorical values to another.
- Useful for collapsing categories or regrouping data for analysis.
- Option to substitute new values for existing field or generate a new field.

vi). Set To Flag 

- Derive flag fields based on categorical values defined for one or more nominal fields.
- Also provide functionality for aggregate keys.

(4) Graphs

- Uses graphs & charts to explore data & gain insights into data types & distributions.
- Check distributions & relationships b/w newly derived fields.

i). Graphboard 

- Offers many different types of graphs in one single node.
- Choose the data fields to explore & then, select a graph from those available for selected data
- Automatically filters out any graph types that would not work with field choices.

ii). Distribution 

- Shows occurrence of symbolic (non-numeric) values in a dataset.
- Show imbalance in data that can be rectified using the Balance node before creating a model.
- Provides color overlay, illustrating distribution of a field's values with each value of specified field.

iii). Histogram 

- Shows occurrence of values for numeric fields
- Used to explore data before manipulation & model building
- frequently used to reveal imbalances in data
- Provides color overlay
- Can use Generate menu to create Balance, Select or Derive Nodes using data in graph

(5) Modeling

- Offers a variety of modeling methods taken from ML, AI & statistics.
- Derive new info from data & develop predictive models
- Each method has certain strengths & is best suited for particular types of problems.

i). CHAID 

- Chi-squared Automatic Interaction Detection
- Classification method builds decision trees by using chi-square statistics to identify optimal splits
- Examines cross-tabulations b/w each input fields & outcome, & tests for significance using a chi-square independence test
- If one of these relations is statistically significant, CHAID selects the most significant one (smallest P value).
- If inputs have more than 2 categories → compared categories will no differences in outcome → merged.

↓ Successively done by joining categories showing least significance.

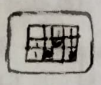
- for nominal input fields, any categories can merge for ordinal set, only contiguous categories can be merged
- CHAID can generate non-binary trees (some splits have more than 2 branches)
- Works for all types of inputs & accepts both case weights & frequency variables.

⑥ Output

- Provides means to obtain info about data & models.
- Provide a mechanism for exporting data in various formats to interface with other software tools.

i). Table 

- Creates a table listing values in your data.
- All fields & values are included to inspect data values or export them in a readable form
- Optionally, highlight records meeting a certain condition.
- Unable to display properly for records > 100 million rows.

ii). Matrix 

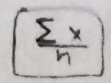
- Creates a table showing relationships b/w fields
- Commonly used to show relationship b/w 2 categorical fields (flag, nominal, ordinal), but it can also be used to show relationships b/w continuous (numeric) fields.

iii). Analysis 

- Evaluate ability of model to generate accurate predictions.
- Perform various comparisons b/w predicted & actual values for one or more model nuggets.
- Compare your model to other predictive models
- A summary of analysis results is automatically added for each model nugget in executed flow.
- Useful with supervised model only.

iv). Data Audit 

- Provides a comprehensive first look at data, presented in an easy-to-read matrix that can be sorted.
- Summary stats, histograms & distribution graphs to gain preliminary understanding of data
- Info about outliers, extremes & missing values.

v). Statistics 

- Basic summary info about numeric fields
- Summary stats for individual fields & correlations b/w fields.

vi). Means 

- Compare means b/w independent groups or b/w pairs of related fields to test whether significant difference exists.
- Comparison in 2 different ways (b/w groups in a field or b/w pair of fields)

⑦ Export

- Produce output mechanism for exporting data

i). Statistics Export

- Outputs data in IBM SPSS Statistics .sav or .zsav format
- Can be read by IBM SPSS Statistics Base & other products.