heig-vd

predictive layer

Better prediction, faster, smarter.

Predictive Analytics Course

Developers Audience

CONTACT
Mohamed Bibimoune
+33 6 70 84 50 80
mohamed.bibimoune@predictivelayer.com

Version 1.0

1.0 /

Agenda

- Part 1: Introduction to Predictive Analytics (Techniques, Algorithms, and Examples)
- Part 2: Practical work (detailed case study)
- Part 3: Practical work (Advanced Techniques/Team competition)

1.0 /

Introduction – Predictive Analytics

Predictive Analytics refers to all techniques, approaches, concepts and processes that aim at leveraging knowledge from data :

- Identify risks and opportunities
- Take better decision at the best moment
- Non bias knowledge extract from data
- Different from statistical exploration

2.0 /

Standard techniques

Descriptive statistics and inferential statistics to analyze produced events :

- Mean, Standard deviation, R2 correlation
- Normal distribution, Binomial distribution, Geometric distribution
- Hypotheses about data, causality, a-priori
- Expert knowledge for modeling, Bayesian acyclic graphs

3.0 /

Machine Learning techniques

Enhance the computational power of new hardware to discover automatically knowledge from data:

- Data driven, empirical knowledge discovery
- Group of algorithms designed for specific tasks (supervised, clustering, reinforcement learning, semi supervised …)
- Powerful and mostly non parametric, with high dimensional space search
- Used for prediction and generalization

4.0 /

Supervised vs Unsupervised

Supervised:

- Binary classes
- Multiple classes
- Multiple labels
- Regression

Unsupervised:

- Clustering
- Frequent item set
- Anomaly detection

5.0 /

Real world examples

- Fraud detection systems
- Marketing recommendation systems
- Predictive maintenance systems
- Resource/Production/Demand forecasting systems
- Autonomous systems

6.0 /

Data acquisition and extraction

- Draw a map of the existing data sources

- Build ETL:

    - Develop connectors to unify technologies
    - Read the data dictionaries (Entity – Association Diagram)
    - Design the new relational scheme
    - Write the new data-mart dictionary
    - Automate the controls and checks on streamed data
    - Automate the extraction process (Batch or real-time)

7.0 /

Data preparation

- Involve analysts and database experts
- Asses the following points for each variables:
  - Periodicity - Frequency
  - Update rate or versioning database states
  - Latency
  - Primary keys and foreign keys

Data cleansing, Data transformation

- Data cleansing must be done in every iteration of the project (Agile methodology)

- Data transformation is a process closely related to modeling:
  - Handle missing values
  - Categorization
  - Count – Average - Smoothing
  - Projection in new representation spaces
  - Binarization
  - Correlations

9.0 /

Data sampling

- Down sampling:
  - Reduce the number of over-represented instances
  - Reduce the number of useless instances
- Oversampling:
  - Add generated artificial instances
  - Overweight existing under-represented instances
  - Duplicate under-represented instances

Predictor and target variables

- Predictor:
  - Difficult to define the usable ones
  - Try to use as much as possible the relevant ones excluding redundancy
- Target:
  - Define the mathematical variable which precisely answer to the question expected
  - Use the right format for encoding the target
  - Ensure the usability of the predictions

11.0 /

Type of prediction

- Classification:
  - Binary classes, Multi-classes, Multi-label
  - Select the right algorithms
  - Select the right evaluation metric in accordance with the usage of the prediction
- Regression:
  - Time-series values
  - Quantified values
  - Select the right algorithms
  - Select the right evaluation metrics

12.0 /

Metrics

Different evaluation metrics depending on the modeling:

- ROC
- LIFT
- AUC
- RMSE
- F-SCORE
- LOG-LOSS
- MAE
- ACC
- AVP
- BEP

13.0 /

Algorithms for predictive Analytics

Some of the principal supervised learning algorithms:

- Naive-Bayes
- Ridge Regression
- Lasso
- Decision Trees
- Nearest neighbor
- Neural Network
- Boosting
- Random Forests
- Support Vector Machines

14.0 /

Test and quality of results

Splitting data for better validation:

- Cross validation
- Random split train/test
- Time based train/test
- Out of sample with different distribution

15.0 /

Part 2 – Practical work

Predict the insurance fees for cars using almost native Python function:

- Dataset of historical insurance fees cars
- Understand the provided data manipulation code
- Understand and explore the data through pandas
- Run the simple validation code to handle the project
- Explore the modelling opportunities
- Find new features and speed-up optimization
- Try to improve the best scores

16.0 /

Part 3 – Practical work – Team job

Build teams to improve the performances of submission for the competition

- 4 teams:
  - Data visualization and understanding
  - Feature engineering
  - Validation protocol
  - Modelling team

- Gather all pieces of codes:
  - Make a submission to evaluate the performance
  - Try to improve the models

- Change the team members:
  - Try other submissions

- Summarize the difficulties encountered and how can they be handled for next projects

Part 3 – Advanced Methods

- Ensemble is the best way to improve predictive analytics
- Bring diversity in your approaches
- Handling high dimensional categorical future is the key point
- Use algorithms robust to noise and missing data
- Produce simple codes that make smart tasks

18.0 /

Conclusion

Think more – Try less

Identify valuable data – Choose the right representation

Ask the right question – Implement the right solution

Properly validate – Always try the worst cases

Be part of competitions – Always try new technologies – new algorithms and new feature engineering

Read a lot about what other developers do predictive layer

Better prediction, faster, smarter.

+33 (0)614 677 082
+33 (0)670 845 080
contact@predictivelayer.com

www.predictivelayer.com

heig-vd