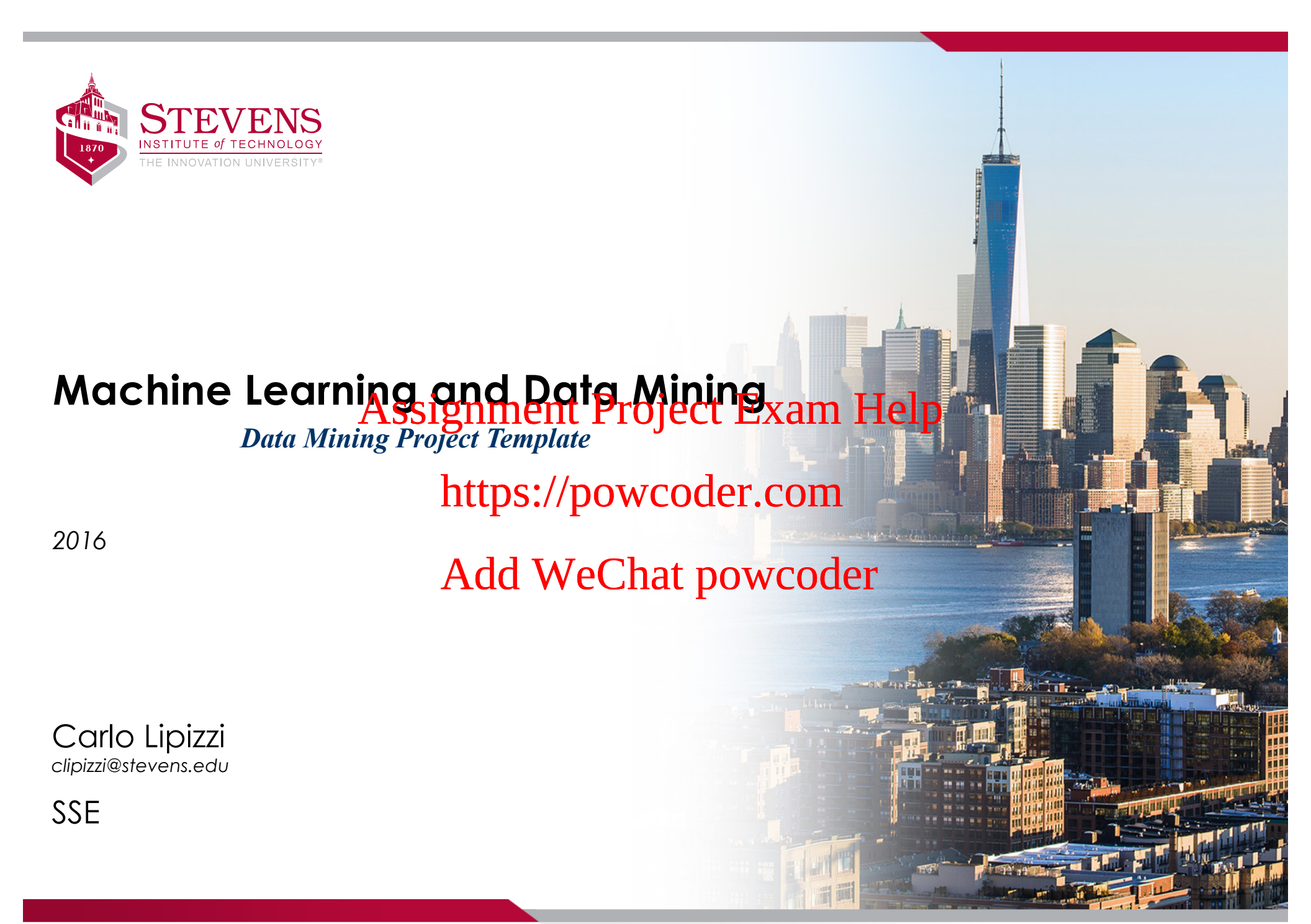


Machine Learning and Data Mining

Data Mining Project Template

2016

Carlo Lipizzi
clipizzi@stevens.edu

SSE

Contents

- Project Goals and Conditions
- CRISP
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- Practical Results – Conclusions
- Attachments

Project Goals and Conditions

- What are the project goals? What is the key question you are required to answer?
- Are there any conditions limiting or somehow defining the project, like limited access to data, data too old, time constrains
- A brief description of the expected results may be added

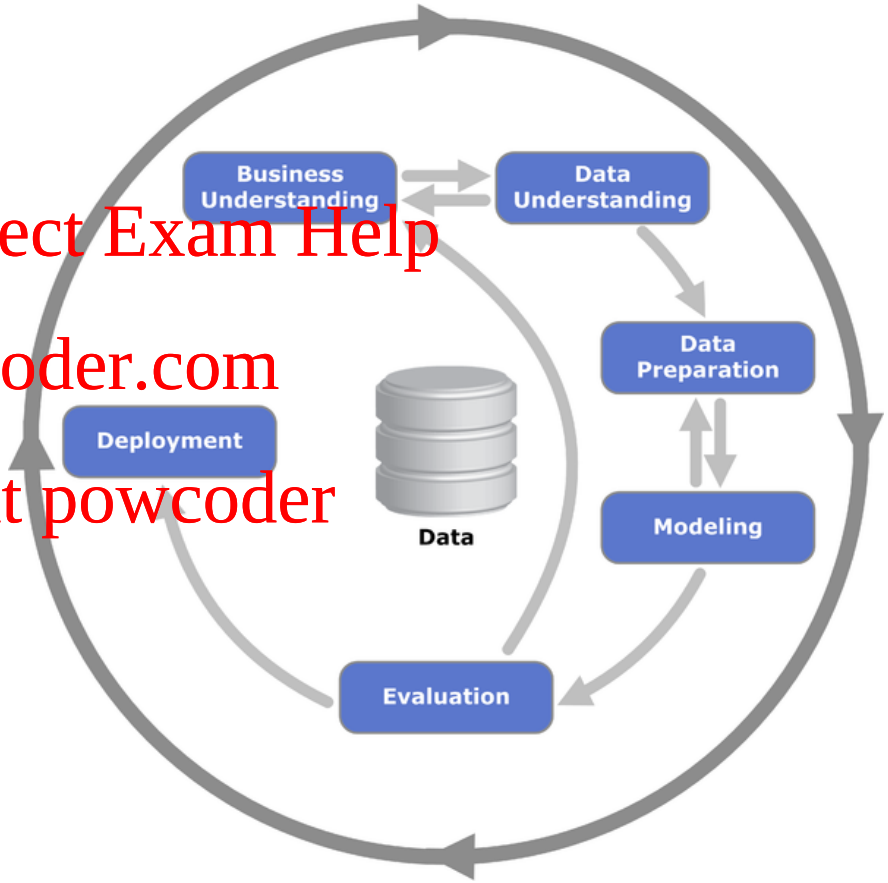
Contents

- Project Goals and Conditions
- CRISP
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- Practical Results – Conclusions
- Attachments

CRISP-DM

Background Info & Definition*

- Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996
- Developed to fit data mining into general business strategy
- Process vendor and tool-neutral
- Non-proprietary and freely available
- Data mining projects follow iterative, adaptive life cycle consisting of 6 phases



*: *from D. Larose – Discovering Knowledge in Data*

Business Understanding

- **Definition***
  - Define business requirements and objectives
  - Translate objectives into data mining problem definition
  - Prepare initial strategy to meet objectives
- **You want to be sure to clearly describe the business needs and the steps to address them**

*: *from D. Larose – Discovering Knowledge in Data*

Data Understanding

- **Definition***
  - Collect data
  - Assess data quality
  - Perform exploratory data analysis (EDA)
- **Overall data description: sources, organization, key characteristics (sensor/human generated, reliable/unreliable source, …)**
- **Here you run all the descriptive statistical tests that make sense for the specific case, describing the different steps and their specific meanings**

*: *from D. Larose – Discovering Knowledge in Data*

Data Preparation

- **Definition***
  - Cleanse, prepare, and transform data set
  - Prepares for modeling in subsequent phases
  - Select cases and variables appropriate for analysis
- **First define the steps you are going to perform (e.g.: if you normalize, why)**
- **Here you perform all the data transformation applicable to the case: missing/miscalculated/misplaced values, outliers, normalization**
- **Describe the final dataset (format, new records number, new variables, …)**

*: *from D. Larose – Discovering Knowledge in Data*

Modeling

- **Definition***
  - Select and apply one or more modeling techniques
  - Calibrate model settings to optimize results
  - If necessary, additional data preparation may be required
- **Explain why you selected a model to an other**
- **Explain the setting parameters you chose (high level description only)**
- **Describe the first results and eventually the adjustments you made**
- **Describe eventual adjustments you made back to the data**
- **Describe final results**

*: *from D. Larose – Discovering Knowledge in Data*

Evaluation

- **Definition***
  - Evaluate one or more models for effectiveness
  - Determine whether defined objectives achieved
  - Make decision regarding data mining results before deploying to field
- **Some models can be evaluated using part of the data you have (supervised learning). In this case, describe the results, using reliable metrics (e.g.: error/confusion matrix)**
- **If the model is unsupervised (no data for testing), evaluate your data using a reliable key performance evaluator (KPI), from outside the perimeter of your data, eventually using your knowledge of the domain**
- **Read the results with business sense and provide your comments**

*: *from D. Larose – Discovering Knowledge in Data*

Deployment

- **Definition***
  - Make use of models created
  - Simple deployment: generate report
  - Complex deployment: implement additional data mining effort in another department
  - In business, customer often carries out deployment based on model
- **If the output is a model to be used in real life, be sure it can be exported in a format that can work in the target environment**
- **If the output is a model that is not going to run in real life (e.g.: proof of concept, demo) produce all the reports that may be necessary to fully explain the model and its value in this specific case**

*: *from D. Larose – Discovering Knowledge in Data*

Contents

- Project Goals and Conditions
- CRISP
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- Practical Results – Conclusions
- Attachments

Conclusions

- This is the final recap: you briefly describe the whole process, from the business need, to the data collected, to the model you built, to the results you obtained
- Describe the advantages in using the model, compared to no model or previous models
- Describe possible limitations of the model and future possible developments

Contents

- Project Goals and Conditions
- CRISP
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- Practical Results – Conclusions
- Attachments

Attachments

- All the tables and graphs will go here
- Add only the outputs that can support the case you described in previous slides
- Outputs have to be either readable (no 1M row table in 1 page)