

YELP DATASCIENCE PROJECT

yfe

January 15, 2017

Contents

1	Business understanding	3
1.1	Background	3
1.2	Business objectives and success criteria	3
1.3	Inventory of resources	3
1.4	Requirements, assumptions and constraints	3
1.5	Risks and contingencies	3
1.6	Terminology	3
1.7	Costs and benefits	3
1.8	Data mining goals and success criteria	3
1.9	Project plan	3
1.10	Initial assessment of tools and techniques	3
2	Data understanding	3
2.1	Initial data collection report	3
2.1.1	Background of data	3
2.1.2	List of data sources with broad area of required data covered by each	3
2.1.3	For each data source, method of acquisition or extraction	3
2.1.4	Problems encountered in data acquisition or extraction	3
2.2	Data description report	3
2.2.1	Each data source described in detail	3
2.2.2	List of tables (may be only one or other database objects)	3
2.2.3	Description of each field including units, codes used, etc.	3
2.3	Data exploration report	3
2.3.1	Background including broad goals of data exploration	3
2.4	Data quality report	4
2.4.1	Background including broad expectations about data quality	4

3	Data preparation	4
3.1	Dataset description report	4
3.1.1	Background including broad goals and plan for preprocessing	4
3.1.2	Rationale for inclusion/exclusion of datasets	4
4	Modeling	5
4.1	Modeling assumption	5
4.2	Test design	5
4.2.1	Background - outlines the modeling undertaken and its relation to the data mining goals	5
4.3	Model description	5
4.3.1	Overview of models produced	5
4.3.2	Summary of conclusions	6
4.4	Model assessment	6
4.4.1	Overview of assessments process and results including any deviations from the plan	6
5	Evaluation	6
5.1	Assessment of data mining results with respect to business success criteria	6
5.2	Review of process	7
5.3	List of possible actions	7
6	Deployment	7
6.1	Deployment plan	7
6.1.1	Summary of deployable results	7
6.1.2	Description of deployment plan	7
6.2	Monitoring and maintenance plan	7
6.2.1	Overview of results deployment and indication of which may require updating (and why)	7
6.2.2	Summary of the results updating process	7
6.3	Final report	7

1 Business understanding

- 1.1 Background
- 1.2 Business objectives and success criteria
- 1.3 Inventory of resources
- 1.4 Requirements, assumptions and constraints
- 1.5 Risks and contingencies
- 1.6 Terminology
- 1.7 Costs and benefits
- 1.8 Data mining goals and success criteria
- 1.9 Project plan
- 1.10 Initial assessment of tools and techniques

2 Data understanding

- 2.1 Initial data collection report
 - 2.1.1 Background of data
 - 2.1.2 List of data sources with broad area of required data covered by each
 - 2.1.3 For each data source, method of acquisition or extraction
 - 2.1.4 Problems encountered in data acquisition or extraction
- 2.2 Data description report
 - 2.2.1 Each data source described in detail
 - 2.2.2 List of tables (may be only one or other database objects)
 - 2.2.3 Description of each field including units, codes used, etc.
- 2.3 Data exploration report
 - 2.3.1 Background including broad goals of data exploration

For each area of exploration undertaken:

- Expected regularities or patterns

- Methods of detection
- Regularities or pattern found, expected and unexpected
- Any other surprises
- Conclusions for data transformation, data cleaning and any other pre-processing
- Conclusions related to data mining goals or business objectives
- Summary of conclusions

2.4 Data quality report

2.4.1 Background including broad expectations about data quality

For each dataset :

- Approach taken to assess data quality
- Results of data assessment
- Summary of data quality conclusions

3 Data preparation

3.1 Dataset description report

3.1.1 Background including broad goals and plan for preprocessing

3.1.2 Rationale for inclusion/exclusion of datasets

For each included dataset:

- Description of the preprocessing, including the actions that were necessary to address any data quality issues
- Detailed description of the resultant dataset, table by table and field by field
- Rationale for inclusion/exclusion of attributes

- Discoveries made during preprocessing and any implications for further work
- Summary and conclusions

4 Modeling

4.1 Modeling assumption

4.2 Test design

4.2.1 Background - outlines the modeling undertaken and its relation to the data mining goals

For each modeling task:

- Broad description of the type of model and the training data to be used
- Explanation of how the model will be tested or assessed
- Description of any data required for testing
- Plan for production of test data if any
- Description of any planned examination of models by domain or data experts
- Summary of test plan

4.3 Model description

4.3.1 Overview of models produced

For each model:

- Type of model and relation to data mining goals
- Parameter settings used to produce the model
- Detailed description of the model and any special features (see p. 66)
- Conclusions regarding patterns in the data (if any);

4.3.2 Summary of conclusions

4.4 Model assessment

4.4.1 Overview of assessments process and results including any deviations from the plan

For each model:

- Detailed assessment of model including measurements such as accuracy and interpretation of behavior
- Any comments on models by domain or data experts
- Summary assessment of model
- Insights into why a certain modeling technique and certain parameter settings led to good/bad results
- Summary assessment of complete model set

5 Evaluation

5.1 Assessment of data mining results with respect to business success criteria

- Review of Business Objectives and Business Success Criteria (which may have changed during and/or as a result of data mining)
- Review of Project Success; has the project achieved the original Business Objectives?
- Are there new business objectives to be addresses later in the project or in new projects?
- Conclusions for future data mining projects

5.2 Review of process

5.3 List of possible actions

6 Deployment

6.1 Deployment plan

6.1.1 Summary of deployable results

6.1.2 Description of deployment plan

6.2 Monitoring and maintenance plan

6.2.1 Overview of results deployment and indication of which may require updating (and why)

For each deployed result:

- Description of how updating will be triggered
- Description of how updating will be performed

6.2.2 Summary of the results updating process

6.3 Final report

- Summary of Business Understanding: background, objectives and success criteria.
- Summary of data mining process.
- Summary of data mining results.
- Summary of results evaluation.
- Summary of deployment and maintenance plans.
- Cost/benefit analysis.
- Conclusions for the business.
- Conclusions for future data mining.