Business Data Mining
Semester 2, 2019

Lecture 3
How to Conduct Business Understanding Phase?

Kim, Yang Sok

Dept. of MIS, Keimyung University

Table of Content

Introduction

Introduction

- **In the Business Understanding phase of a data-mining project, before you approach data or tools, you define what you're out to accomplish and define the reasons for wanting to achieve this goal.**

- **The business understanding phase includes four tasks (primary activities, each of which may involve several smaller parts).**
  - Identifying your business goals
  - Assessing your situation
  - Defining your data-mining goals
  - Producing your project plan

Tasks & Outputs

- **The first thing you must do in any project is to find out exactly what you're trying to accomplish!**

- **You must start with a clear understanding of**
  - **Problem**
    - What does your management want to address
  - **Business goals**
  - **Constraints**
    - limitations on what you may do, the kinds of solutions that can be used, when the work must be completed, and so on
  - **Impact**
    - how the problem and possible solutions fit in with the business

- **Background:**
  - Explain the business situation that drives the project with a few paragraphs.

- **Business goals:**
  - Define what your organization intends to accomplish with the project.
  - e.g., Increase sales from a holiday ad campaign by 10 percent over year.

- **Business success criteria:**
  - Define how the results will be measured.
  - Try to get clearly defined quantitative success criteria.
  - If you must use subjective criteria (hint: terms like gain insight or get a handle on imply subjective criteria), at least get agreement on exactly who will judge whether or not those criteria have been fulfilled.

- This is where you get into more detail on the issues associated with your business goals.

- Now you will go deeper into fact-finding, building out a much fleshier explanation of the issues outlined in the business goals task.

Assessing Situation - Outputs

- **Inventory of resources:**
  - A list of all resources available for the project. These may include people (not just data miners, but also those with expert knowledge of the business problem, data managers, technical support, and others), data, hardware, and software.
- **Requirements, assumptions, and constraints:**
  - Requirements will include a schedule for completion, legal and security obligations, and requirements for acceptable finished work. This is the point to verify that you'll have access to appropriate data!
- **Risks and contingencies:**
  - Identify causes that could delay completion of the project, and prepare a contingency plan for each of them. For example, if an Internet outage in your office could pose a problem, perhaps your contingency could be to work at another office until the outage has ended.
- **Terminology:**
  - Create a list of business terms and data-mining terms that are relevant to your project and write them down in a glossary with definitions (and perhaps examples), so that everyone involved in the project can have a common understanding of those terms.
- **Costs and benefits:**
  - Prepare a cost-benefit analysis for the project. Try to state all costs and benefits. If the benefits don't significantly exceed the costs, stop and reconsider this analysis and your project.

- **Reaching the business goal often requires action from many people, not just the data miner.**

- **So now, you must define your little part within the bigger picture.**

- **If the business goal is to reduce customer attrition, for example, your data-mining goals might be to identify attrition rates for several customer segments, and develop models to predict which customers are at greatest risk.**

- **Data-mining goals:**
  - Define **data-mining deliverables**, such as **models, reports, presentations, and processed datasets.**

- **Data-mining success criteria:**
  - Define the **data-mining technical success criteria** necessary to support the business success criteria.
  - Try to define these in **quantitative terms** (such as model accuracy or predictive improvement compared to an existing method).
  - If the criteria must be **qualitative**, identify the person who makes the assessment.

- Describe the intended plan for achieving the data mining goals and thereby achieving the business goals.

- Your plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques.

- **Project plan:**
  - Outline your step-by-step action plan for the project.
  - Expand the outline with a schedule for completion of each step, required resources, inputs (such as data or a meeting with a subject matter expert), and outputs (such as cleaned data, a model, or a report) for each step, and dependencies (steps that can't begin until this step is completed).
  - **Explicitly state that certain steps must be repeated** (for example, modeling and evaluation usually call for several back-and-forth repetitions).

- **Initial assessment of tools and techniques:**
  - Identify the required capabilities for meeting your data-mining goals and assess the tools and resources that you have.
  - If something is missing, you have to address that concern very early in the process.

Data Mining Success Criteria

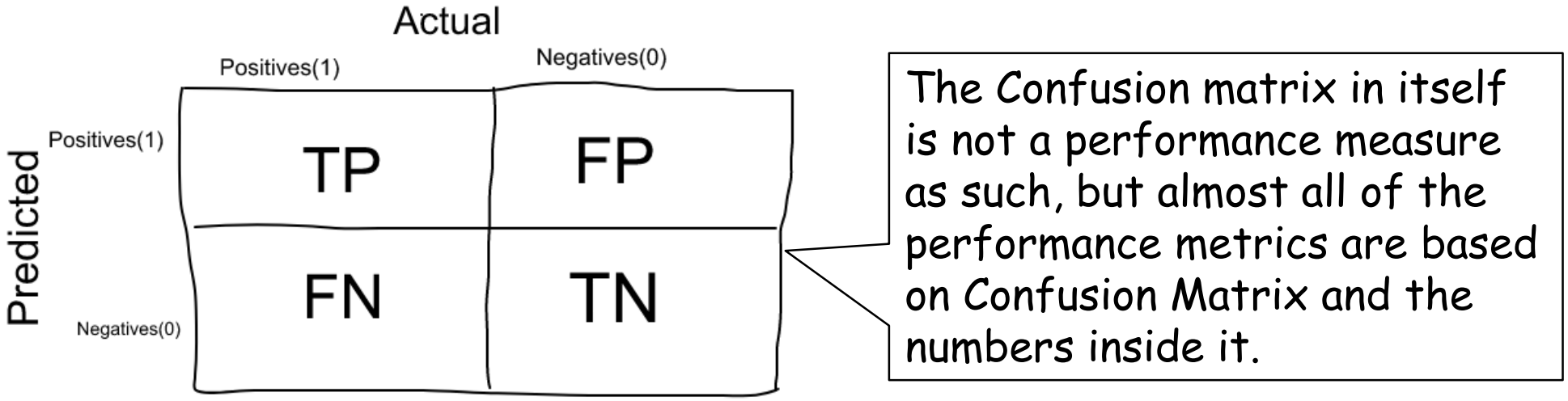
https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Model Performance Measures

Measures

- Performance measures of the model depends on the data mining task
- In this lecture, we focuses on the following three problems – classification and regression.
- Note that classification and regression have correct answers, while clustering has no correct answer.
- This lecture only covers the concept of the performance measures and will practice the following lectures.

- **Confusion matrix**
  - The Confusion matrix is one of the most intuitive and easiest (unless of course, you are not confused)metrics used for finding the correctness and accuracy of the model. It is used for classification problem where the output can be of two or more types of classes.
  - The confusion matrix, is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Our Actual classifications are columns and Predicted ones are Rows.



***1**: When a person is having cancer **0:** When a person is NOT having cancer.*

Model Performance Measures

Classification Metrics

- **Terms associated with Confusion matrix:**
  - **True Positives (TP):** True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True).
    - Ex: The case where a person is actually having cancer(1) and the model classifying his case as cancer(1) comes under True positive.
  - **True Negatives (TN):** True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False).
    - Ex: The case where a person NOT having cancer and the model classifying his case as Not cancer comes under True Negatives.
  - **False Positives (FP)**: False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1)
    - Ex: A person NOT having cancer and the model classifying his case as cancer comes under False Positives.
  - **False Negatives (FN)**: False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)
    - Ex: A person having cancer and the model classifying his case as No-cancer comes under False Negatives.

The ideal scenario that we all want is that the model should give 0 False Positives and 0 False Negatives. But that's not the case in real life as any model will NOT be 100% accurate most of the times.

- **When to minimize what? (1)**
  - We know that there will be some error associated with every model that we use for predicting the true class of the target variable. This will result in **False Positives and False Negatives**(i.e Model classifying things incorrectly as compared to the actual class).
  - There's no hard rule that says what should be minimized in all the situations. It purely depends on the business needs and the context of the problem you are trying to solve.
  - Based on that, we might want to minimize either False Positives or False negatives.

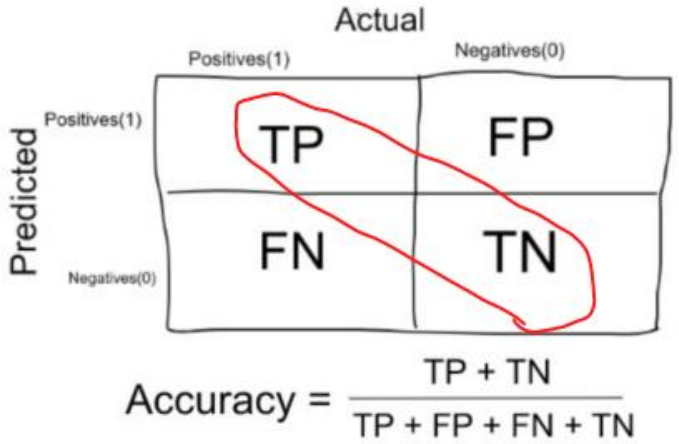
- **When to minimize what? (2)**
  - **Minimizing False Negatives:** Let's say in our cancer detection problem example, out of 100 people, only 5 people have cancer. In this case, we want to correctly classify all the cancerous patients as even a very BAD model(Predicting everyone as NON-Cancerous) will give us a 95% accuracy(will come to what accuracy is).
  - But, in order to capture all cancer cases, we might end up making a classification when the person actually NOT having cancer is classified as Cancerous.
  - This might be okay as it is less dangerous than NOT identifying/capturing a cancerous patient since we will anyway send the cancer cases for further examination and reports.
  - But missing a cancer patient will be a huge mistake as no further examination will be done on them.

- **When to minimize what? (3)**
  - **Minimizing False Positives:** For better understanding of False Positives, let's use a different example where the model classifies whether an email is spam or not
  - Let's say that you are expecting an important email like hearing back from a recruiter or awaiting an admit letter from a university. Let's assign a label to the target variable and say,1: "Email is a spam" and 0:"Email is not a spam"
  - Suppose the Model classifies that important email that you are desperately waiting for, as Spam(case of False positive). Now, in this situation, this is pretty bad than classifying a spam email as important or not spam since in that case, we can still go ahead and manually delete it and it's not a pain if it happens once a while.
  - So in case of Spam email classification, minimizing False positives is more important than False Negatives.

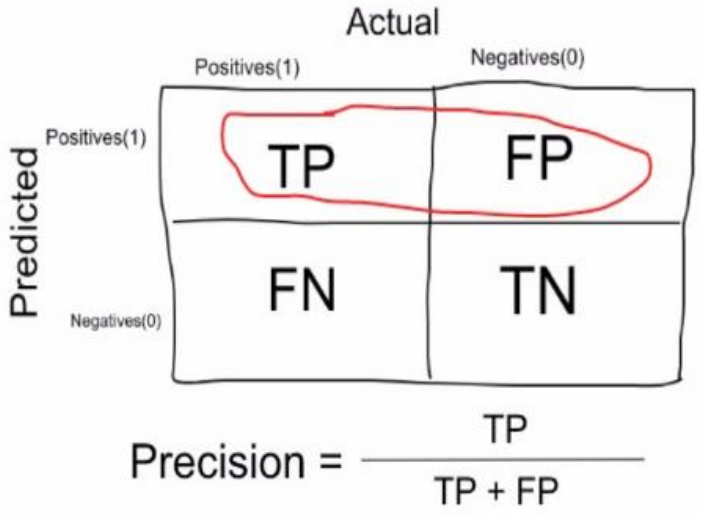
- **Accuracy**
  - This is the number of correct predictions made by the model over all kinds predictions made.

  - **When to use Accuracy:**
    - Accuracy is a good measure when the target variable classes in the data are nearly balanced. Ex:60% classes in our fruits images data are apple and 40% are oranges. A model which predicts whether a new image is Apple or an Orange, 97% of times correctly is a very good measure in this example.

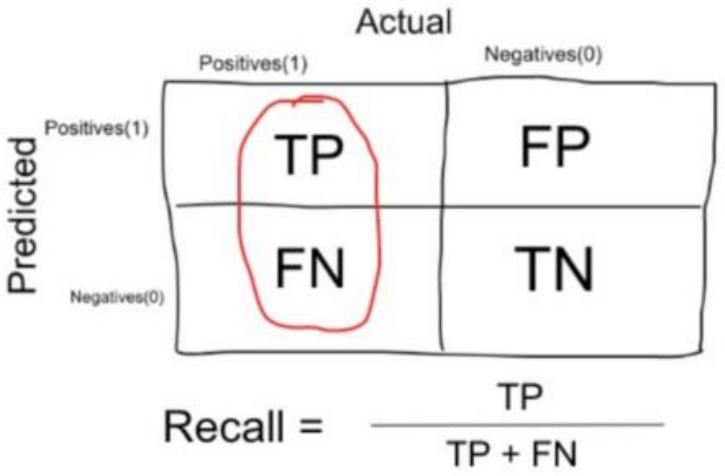
  - **When NOT to use Accuracy:**
    - Accuracy should NEVER be used as a measure when the target variable classes in the data are a majority of one class.



- **Precision**
  - This is the number of correct positive predictions made by the model over all positive predictions made.
  - The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

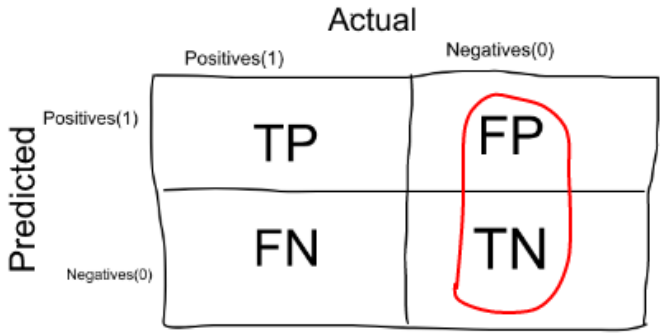


- **Recall**
  - Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer.
  - The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP.
  - Note: FN is included because the Person actually had a cancer even though the model predicted otherwise.



- **When to use Precision and When to use Recall?:**
  - It is clear that recall gives us information about a classifier's performance with respect to false negatives (how many did we miss), while precision gives us information about its performance with respect to false positives(how many did we caught).
    - Precision is about being precise. So even if we managed to capture only one cancer case, and we captured it correctly, then we are 100% precise.
    - Recall is not so much about capturing cases correctly but more about capturing all cases that have "cancer" with the answer as "cancer". So if we simply always say every case as "cancer", we have 100% recall.
  - So basically if we want to focus more on minimizing False Negatives, we would want our Recall to be as close to 100% as possible without precision being too bad and if we want to focus on minimizing False positives, then our focus should be to make Precision as close to 100% as possible.

- **Specificity**
  - Specificity is a measure that tells us what proportion of patients that did NOT have cancer, were predicted by the model as non-cancerous.
  - The actual negatives (People actually NOT having cancer are FP and TN) and the people diagnosed by us not having cancer are TN. (Note: FP is included because the Person did NOT actually have cancer even though the model predicted otherwise).
  - Specificity is the exact opposite of Recall.



$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **F1 Score**
  - **F1 Score** is a single score that kind of represents both Precision(P) and Recall(R).
  - F1 Score = Harmonic Mean(Precision, Recall)
    = 2 * Precision * Recall / (Precision + Recall)
  - So if one number is really small between precision and recall, the F1 Score kind of raises a flag and is more closer to the smaller number than the bigger one, giving the model an appropriate score rather than just an arithmetic mean.

- **RMSE (Root Mean Square Error)**
  - It represents the sample standard deviation of the differences between predicted values and observed values (called residuals). Mathematically, it is calculated using this formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Source: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1

- **MAE**
  - MAE is the average of the absolute difference between the predicted values and observed value.

$$MAE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} ||}$$

  - The MAE is a linear score which means that all the individual differences are weighted equally in the average.

- **So which one should you choose and why?**
  - It is easy to understand and interpret MAE because it directly takes the average of offsets whereas RMSE penalizes the higher difference more than MAE.
  - Generally, RMSE will be higher than or equal to MAE. The only case where it equals MAE is when all the differences are equal or zero.
  - However, even after being more complex and biased towards higher deviation, RMSE is still the default metric of many models because loss function defined in terms of RMSE is smoothly differentiable and makes it easier to perform mathematical operations.

Exercise 1
Write a Data Mining Project Plan

- **Download "Census.zip" from "Teaching & Learning System (교수학습시스템)"**
- **Unzip the file**
- **Read "census_data_description.txt" and guess the followings**
  - What is the goal of the project? Or How can we use this dataset?
  - How they prepared dataset for the project?
  - What type of project they are doing (e.g., prediction, classification, clustering, basket analysis, etc.)?
  - **What measures can be used to evaluate the model performance?**
- **Write a data mining project plan based on your observations**

Exercise 2
Calculate Classification Performance

- **This shows classification results by a model, calculate "Accuracy", "Precision", "Recall", "Specificity" and "F-Measure".**

		Actual	
		Positive	Negative
Predicted	Positive	75	60
	Negative	25	40

Exercise 3
Calculate Regression Performance

Exercise

Exercise 2 Calculate Regression Performance

Age	prediction(Age)
12.500	11.730
11.500	10.827
7.500	8.156
7.500	9.222
10.500	11.004
10.500	12.015
10.500	10.571
10.500	10.993
7.500	8.428
11.500	11.082
14.500	11.328
8.500	9.031
8.500	10.808
8.500	8.919
8.500	8.396
7.500	9.475
11.500	13.794
10.500	10.667
10.500	10.513
10.500	11.938

- **The left figure shows regression results of "Age" prediction. Calculate RMSE and MAE.**

Conclusion

Conclusion

- In Business Understanding phase, it is necessary to identify business goals, assessing situation, defining data mining goals and finally produce project plan.
- A thorough understanding of business is a key condition for the success of a data mining project.
- Otherwise, it makes it hard to get data mining results and consumes resources in vain.

QUESTIONS?