Purwadhika
Digital Technology School

Data Science

Final Project Guideline

JC Data Science – AH BSD-JKT

Guideline Introduction

1. Data : We have provided some dataset with certain level of difficulty
2. Problem : You can formulate a problem based on dataset and problem you get

To DOs Content

To DOs

1. Jupyter Notebook
2. Slide Presentation

Notebook Content

- Problem Statement
- Data Understanding
- Exploratory Data Analysis
- Preprocessing
- Methodology (Modeling/Analysis)
- Conclusion and Recommendation

Slide Presentation Content

- Profile Group Intro
- Problem Formulation
- Data Understanding
- Findings and Solution
- Conclusion and Recommendation

Guideline Detail

Notebook Content

- Problem Statement
- Data Understanding
- Exploratory Data Analysis
- Preprocessing
- Methodology (Modeling/Analysis)
- Conclusion and Recommendation

Notebook Content

Problem Statement

Business Problem Statement

Problem Statement for Machine Learning :

- What is being predicted and the impact ?
- Example : How to predict **house of a price** so **we can minimize overpricing and underpricing phenomenon** ?
- General Formula : How to predict < value > so <goals> ?
  - value : House Price
  - goals : minimize overpricing and underpricing

Business Problem Statement

Problem Statement for Analytics :

- Example : What customer should we target in order to increase revenue ?
- General Formula : What customer should we < action > in order to <goals> ?
    - action : targeting
    - goals : increase revenue

Notebook Content

Data Understanding

Data

Unit Analysis

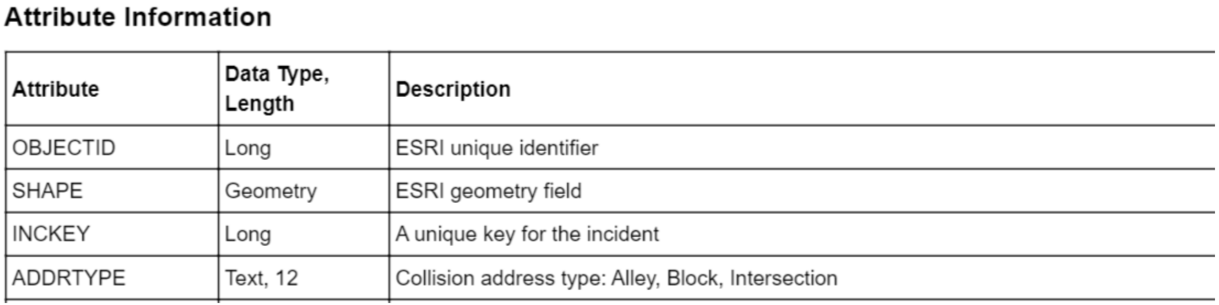
- Represent Each Row

Attributes :

- Data Type (string, numerical, binary, date etc)
- Attributes Description

Unit Analysis in each cases :

- Accident : Accident ID
- One row represent an accident



Attribute Information

Attribute	Data Type, Length	Description
OBJECTID	Long	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
ADDRTYPE	Text, 12	Collision address type: Alley, Block, Intersection

Data (Specific for Machine Learning)

Unit Analysis

- Represent Each Row

Label :

- The label should come after the feature, because it’s predictive (**if any**)

Features :

- The features should come before the label, because it’s predictive
- The features should be available when prediction needed.
- A feature should be a phenomenon that related with the problem

Unit Analysis in each cases :

- Credit scoring : Customer ID
- Fraud : Transaction ID
- House Price : House ID

Notebook Content

Exploratory Data Analysis (EDA)

Exploratory Data Analysis

Objective: To understand data distribution and condition for preprocessing stage

Elements:

1. Data Distribution Plot (Boxplot, QQplot, Histogram)
2. Data Correlation (Nominal and Ratio scale correlation)
3. Data Cardinalities (Count unique for categorical feature)
4. Identify missing value, outlier, anomaly, duplicates, etc

Notebook Content

Preprocessing

Preprocessing

In this part, you should prepare your data for Analysis:

1. Fill missing value
2. Casting your data type
3. Inconsistent Variable
4. Remove data duplication
5. Feature Engineering, etc

Notebook Content

Methodology (Data Analytics)

Analysis Guideline

1. Analysis Paradigm
2. Analysis Structure
3. Expected Analysis Results

Analysis Paradigm

In a way of solving problem, aside from utilizing machine learning, we can also applying data analysis principle that we learned before. The paradigm of analysis can be divided into two types:

1. Descriptive Analysis
   Answering a problem by tell the story about the data descriptively, using statistical approach or visualization.
2. Inferential Analysis
   Answering a problem by hypothesis proof by statistical test (generalization)

Analysis Paradigm

In a way of using those approaches, there is “do” and “dont” guideline about descriptive analysis:

DOs	DONTs
1 graph contain only 1 narration/story	Tell more than 1 narration in 1 graph.
Use Pie Chart to compare categorical data with <= 4 unique values. Use Bar Chart for > 4 unique values.	Use pie chart for >= 4 unique values, or bar chart for < 4 unique values.
Strengthen the findings by adding a theory (from a specific domain expertise) that is related and valid to the case	Use a theory that is not related/valid to the findings

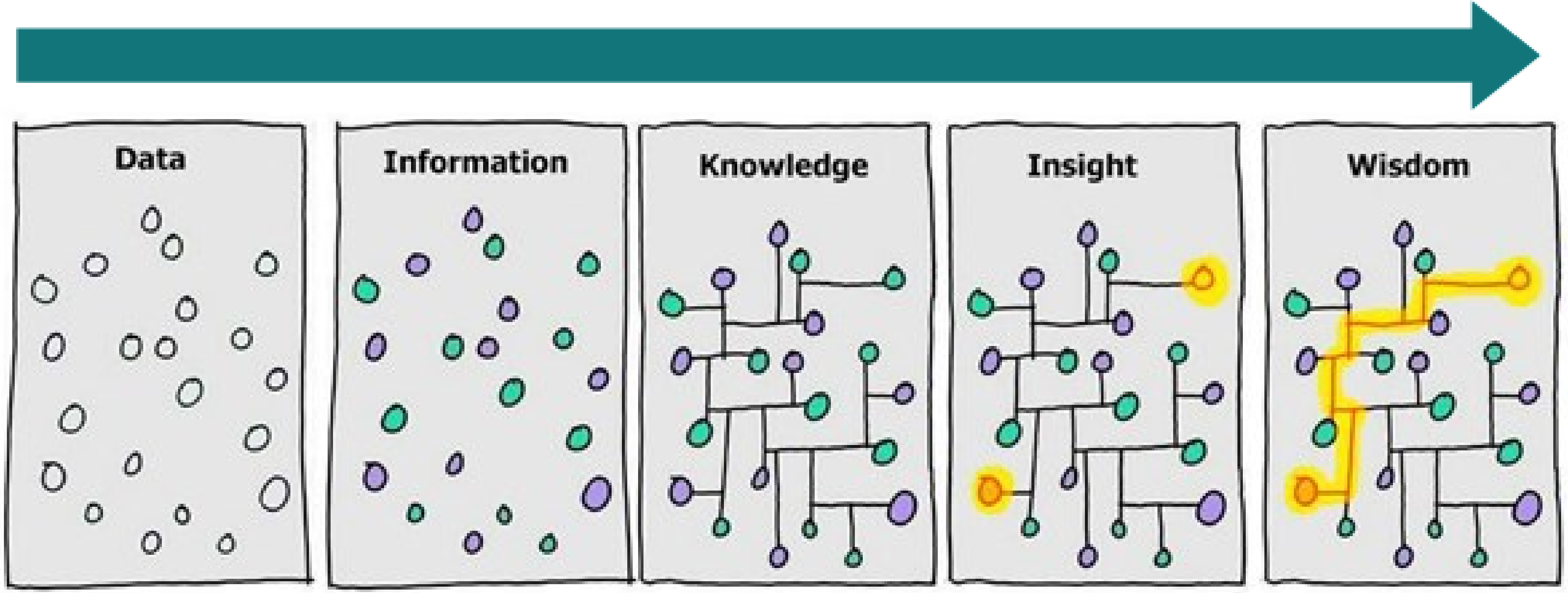
Analysis Paradigm

While the “DOs” and “DONTs” on inferential analysis is:

DOs	DONTs
Always use suitable the statistical test based on research objective and data characteristic.	Use statistical test that is irrelevant either to the problem objective or data characteristic
Wise on determining the level of significance, not too high or not too low. Based on data risk.	Generalizing the level of significance of any statistical test for all types of data.
Wise on sampling selection and sampling method, and not doing cherry picking.	Randomly select the data without any rule/consideration.

Analysis Paradigm

We can also use this analysis framework:



Objective: Connect the dots to find a significance**. The more dimension/variable used, the stronger the insight could be.**

Purwadhika
Digital Technology School

Analysis Paradigm

The more variable are connected, the more powerful insight that we get:

Level/Stage	Description	Example
Data	Show only a raw data frame	The first 5 columns/row in a data frame (df.head())
Information	Summary statistics on a variable	Jumlah pelanggan wanita sebanyak 50 orang
Knowledge	Summary statistics on two variables	Jumlah transaksi oleh pelanggan wanita sebesar 100 transaksi
Insight	Summary statistics on three or more variables, but this is not align with business problem	Jumlah transaksi berdasarkan jenis kelamin dan asal daerah. “Kita mengetahui bahwa transaksi tertinggi berasal dari daerah X dan jenis kelamin Y”
Wisdom	Summary statistics on three or more variables, and align with business problem + relevant action items	Jumlah transaksi berdasarkan jenis kelamin, asal daerah, hari transaksi, dan tipe produk. “Untuk memaksimalkan revenue, kita perlu memberi promo di hari W untuk jenis barang X, sebab transaksi tertinggi terjadi disana oleh jenis kelamin Y, puncaknya pada pelanggan asal daerah Z”

Purwadhika
Digital Technology School

Analysis Structure

As an analyst, structure and analysis quality is very important so can convince our stakeholder to take action. Here is the thing to note about making insight structure:

1. Contains 5W-1H
2. Insight Narration
3. Quantitative Principle
4. Actionable Insight

Analysis Structure

1. Contains 5W-1H
   Insight **can answer 5W-1H** (What, When, Who, Where, Why, and How)

   Example:

   [No] → "Kita memiliki 23 variabel dengan 10 ribu row data"
   (This is not answering any type of problem)

   [**Yes**] → "Pelanggan Pria memiliki jumlah transaksi tertinggi pada pemegang kartu Biru dibandingkan Wanita"
   (Answering the WHO and WHAT)

Analysis Structure

2. Insight Narration
   Insight has a narration/story, so this can generate new perception about our point of view.

   Example:

   [No] → "Rata-rata transaksi pelanggan berkartu biru = 30, merah = 20"
   (No narration, just a raw statistical findings)

   [**Yes**] → "Mayoritas pelanggan Pria yang memiliki kartu biru mempunyai transaksi sebesar 30, dimana nilai ini 10 lebih tinggi dari yang memiliki kartu berwarna merah"
   (There is a narration, it has storytelling point)

Analysis Structure

3. Quantitative Principle

   Insight should displays number (quantitative) to compare among objects

   Example:

   [No] → "Pelanggan Wanita lebih banyak dari Pria pada Kepemilikan Kartu Biru"
   (Context "more" is not clear, there is no number)

   [**Yes**] → "Pelanggan Wanita lebih banyak sebesar 30% dari Pria pada Kepemilikan Kartu Biru"
   (Context "more" is clear, about 30%)

Analysis Structure

4. Actionable Insight
   Insight contains information, recommendation, and potential impact that is relevant to solve the problem.

   Example:

   [Yes] → "Pelanggan Churn total transaksinya selalu kurang dari 30" (Insight is good, but there is no recommendation)

   [**Recommended**] → "Pelanggan Churn total transaksinya dibawah 30, oleh karena itu kita perlu meningkatkan transaksi pelanggan supaya selalu diatas 30 agar peluang churn mereka berkurang sebesar 50%" (Insight is good, also the recommendation + potential impact are relevant to the problem case)

Expected Analysis Result

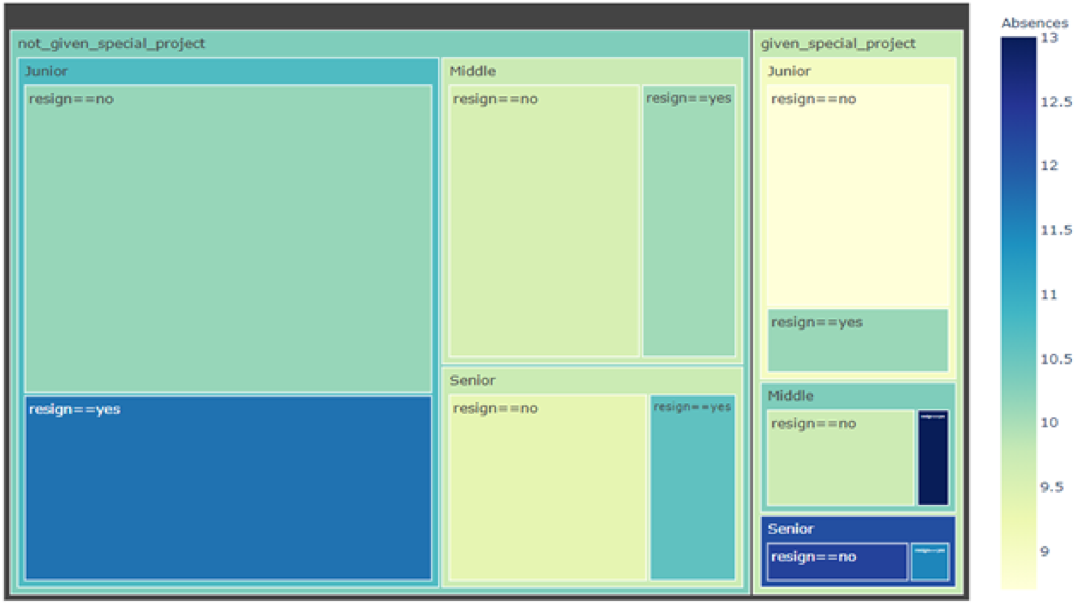
In producing a good standardization, the following is an example of the expected analysis results.

1. Analysis through data visualization
2. Analysis through data storytelling

Expected Analysis Result

1. Analysis through data visualization

Employee Attrition by Special Project Involvement, Employment Level and Absences Behavior

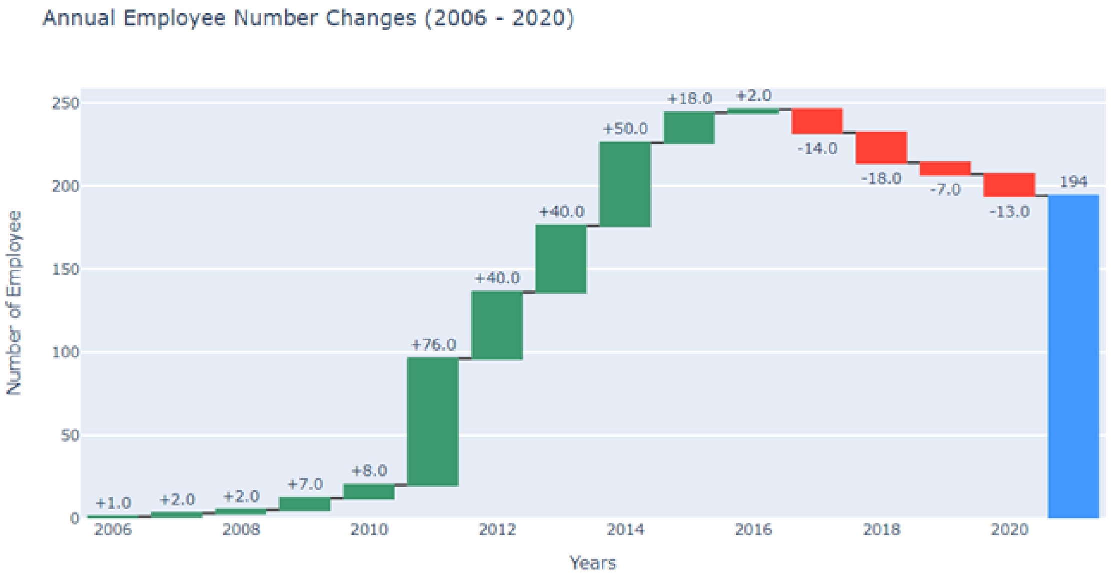


(Membangun persepsi)

Terdapat korelasi kuat antara pemberian/keterlibatan *special project* dengan jumlah absensi. Pegawai yang diberi *special project* memiliki jumlah absensi lebih rendah 1-2 poin dibandingkan yang tidak diberikan *special project*.

Expected Analysis Result

1. Analysis through data visualization

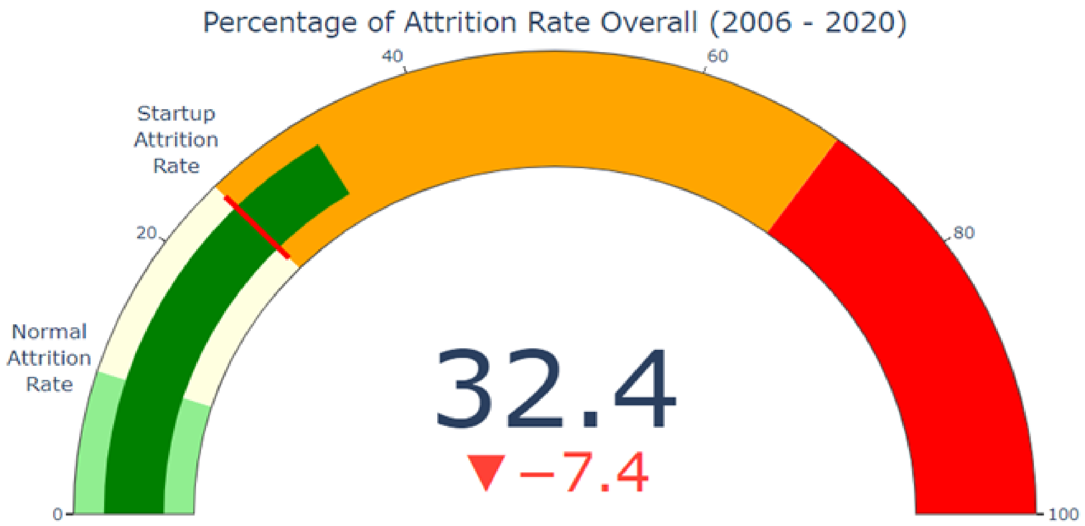


(Membangun narasi)

Sejak awal, dinamika kepegawaian selalu terjadi dari tahun ke tahun. Jumlah pegawai sejak tahun 2006 selalu meningkat yang puncaknya di tahun 2011 sebanyak 76 pegawai. Sedangkan pada 4 tahun terakhir, dari 2017 hingga 2020, jumlah pegawai mengalami penurunan. Di tahun 2021, jumlah pegawai yang tersisa sebanyak 194 pegawai.

Expected Analysis Result

1. Analysis through data visualization



*Normal Attrition Rate: 10% (Deloitte Indonesia, 2019)
*Startup Attrition Rate: 25% (Linkedin, 2018)

(**Mengutip Teori untuk menguatkan Insight**)

Dilansir dari Deloitte Indonesia (2019), normalnya attrition rate sebuah perusahaan yaitu 10%, namun akan berada disekitar angka 25% apabila itu adalah perusahaan startup (Linkedin, 2018)

Perusahaan ini memiliki Attrition Rate sebesar 32.4%, sehingga perusahaan ini masuk dalam kategori perusahaan startup yang dirintis sejak 2006.

Purwadhika
Digital Technology School

Expected Analysis Result

2. Analysis through data storytelling

(Only insight, no action. Score = 50 - 70)

“User yang churn itu lebih banyak yang berpendidikan High School dimana diantara mereka itu kebanyakan pendapatan mereka dibawah $40K per tahun sebanyak 20%”

Expected Analysis Result

2. Analysis through data storytelling

(Good insight, but the recommendation is not relevant. Score = 70 - 80)

“Pelanggan Pria lebih banyak churn pada rentang umur diatas 45 tahun dibandingkan Wanita sebesar 60%. Supaya kita mendapatkan pelanggan yang setia, maka kita perlu menambah pelanggan wanita yang berumur di atas 45 tahun”

Expected Analysis Result

2. Analysis through data storytelling

(Insight, Recommendation, and Action item are good and relevant to the case. Score = 80 - 100)

“Pelanggan yang churn mayoritas memiliki total transaksi dibawah 3 setiap bulannya. Untuk menanggulangi itu, kita sarankan tim bisnis untuk memberikan promo untuk mendongkrak transaksi peserta yang memiliki kurang dari 3 setiap bulannya melalui voucher X guna mengurangi peluang churn sebesar 30%”

Notebook Content

Methodology (Machine Learning)

Data Illustration : Credit Scoring

Features

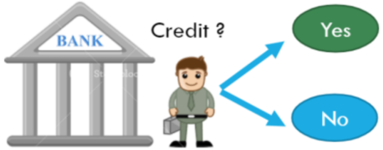
Label

Cust ID	Age	Gender	...	Edu	Balance	...	Income	...	...	Default
										BAD
										BAD
										GOOD
										...
										...
										GOOD

About identity/demographic

About their transaction

etc



Unit Analysis : Customer ID

Data Illustration : Fraudulent Credit Card

Features

Label

Trans ID	Cust ID	Amont	Location	...	Time	Fraud
A102-121	A102					Yes
A102-123	A102					Yes
A102-124	A102					No
B203-111	B203					...
C204-202	C204					...
C204-232	C204					Yes

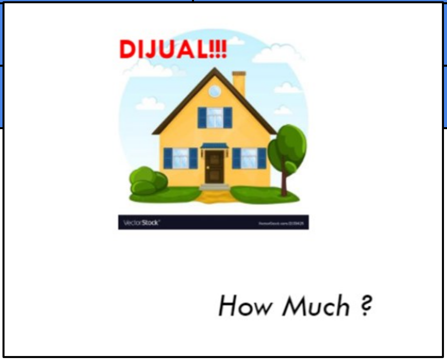
Unit Analysis : Transaction ID

Data Illustration : House Price

Features

Label

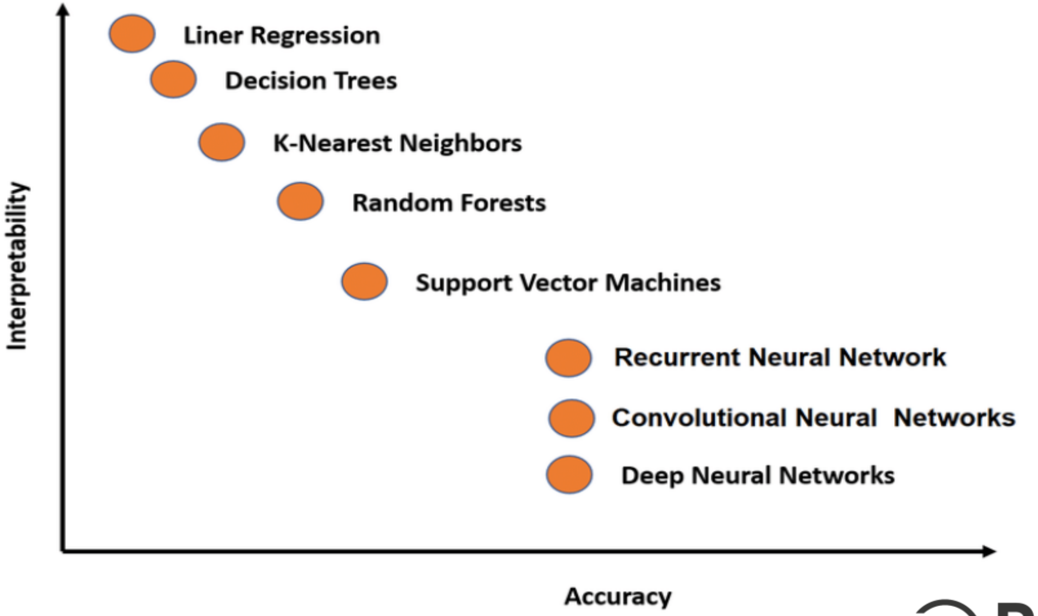
House ID	Number of Bedrooms	Number of Rooms	...	Garage (yes/no)	Price (IDR)
					300,000,000
					400,000,000
					250,000,000
					...
					...
					1,000,000,000



Unit Analysis : House ID

Modeling Approach

What kind of model you need to use to solve your Model



Evaluation For ML Models

How To Optimize Your Model ?

- Should I Use Hyperparameter Tuning ?
- How do you combine preprocessing and modeling (Pipeline) ?

What method you use to evaluate your model ?

- Cross Validation
- Strat. Cross Validation, etc

What is the most suitable metrics you use to evaluate your model ?

- Your metrics should relate to your business problem
- Classification : accuracy, recall, precision, etc
- Regression : MSE, MSLE, etc
- Clustering : Silhouette Score, F-statistics, Average Sum Square Error, etc

Notebook Content

Conclusion and Recommendation

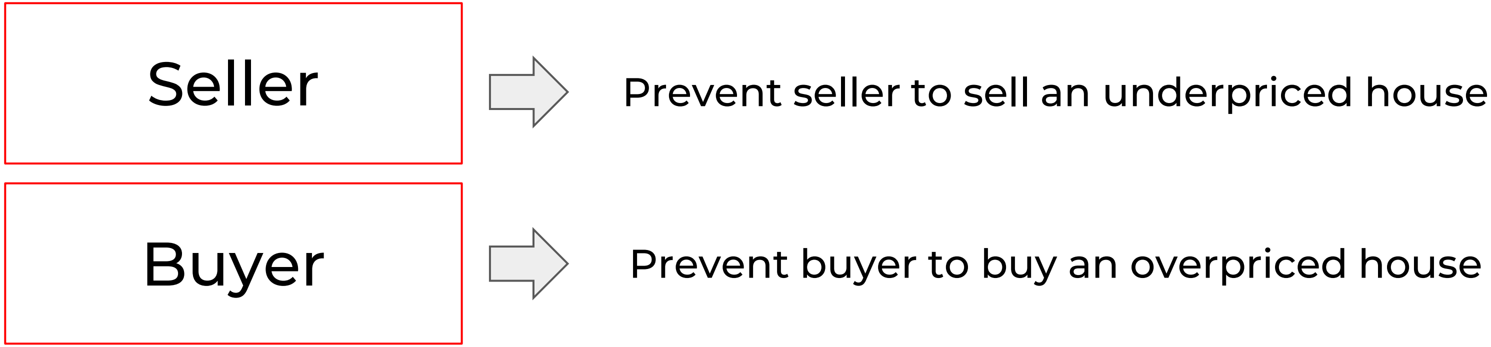
Recommendation: Credit Scoring

Allocate loan to the new applicant based on predicted default risk. Lender can reject new applicant when the risk is above certain tolerance.

Recommendation: Fraudulent Credit Card

Follow up suspected account and do **further investigation** to the transaction detected as fraud so we can actually **recognize fraudulent credit card transaction** and **prevent fraud** from happening. This is an action to minimize phenomenon customer charged for items that they did not purchase.

Recommendation: House Price



Seller → Prevent seller to sell an underpriced house

Buyer → Prevent buyer to buy an overpriced house

Value

Value for each cases :

- credit scoring : profit increase
- fraudulent credit cards : customer satisfaction
- house price prediction : price efficiency

Slide Presentation

Tutorial

Slide Presentation Tutorial

Building an effective presentation slide, here is some tips to reach better performance:

1. Use only 3 main colors (exclude images), the combination should represent these principles: **contrast, gradient, strong**
2. Do not use Times New Roman font, please use tidy font design such as ‘**Montserrat**’ or ‘**Lato**’ as a semi-formal font design.
3. Use at least 15’ for font size for body text to make sure people can see your materials
4. For better design, you can refer/reuse template from slidesgo.com
5. Consider the proportion of image and text inside one slide to be at least 50:50 proportion for better visualization

Purwadhika
Digital Technology School

Thank You