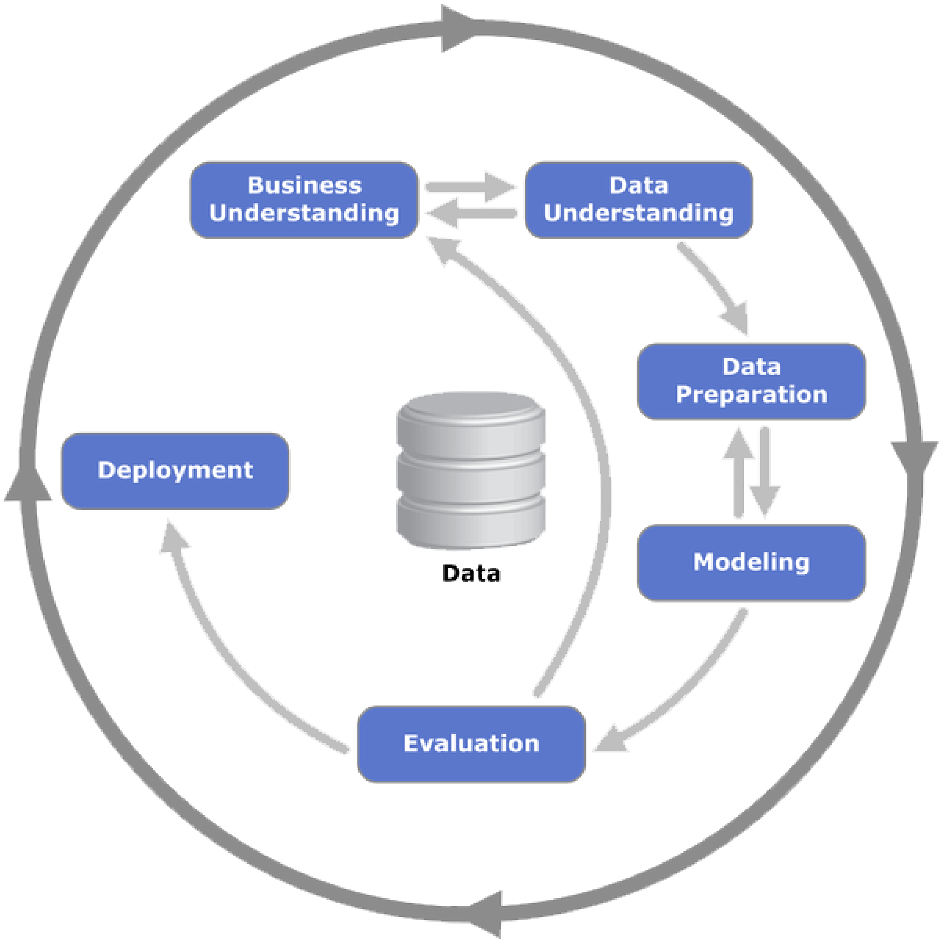
Case Study : Predictive Lead Scoring

Team Algoritma

11/10/2020

Contents

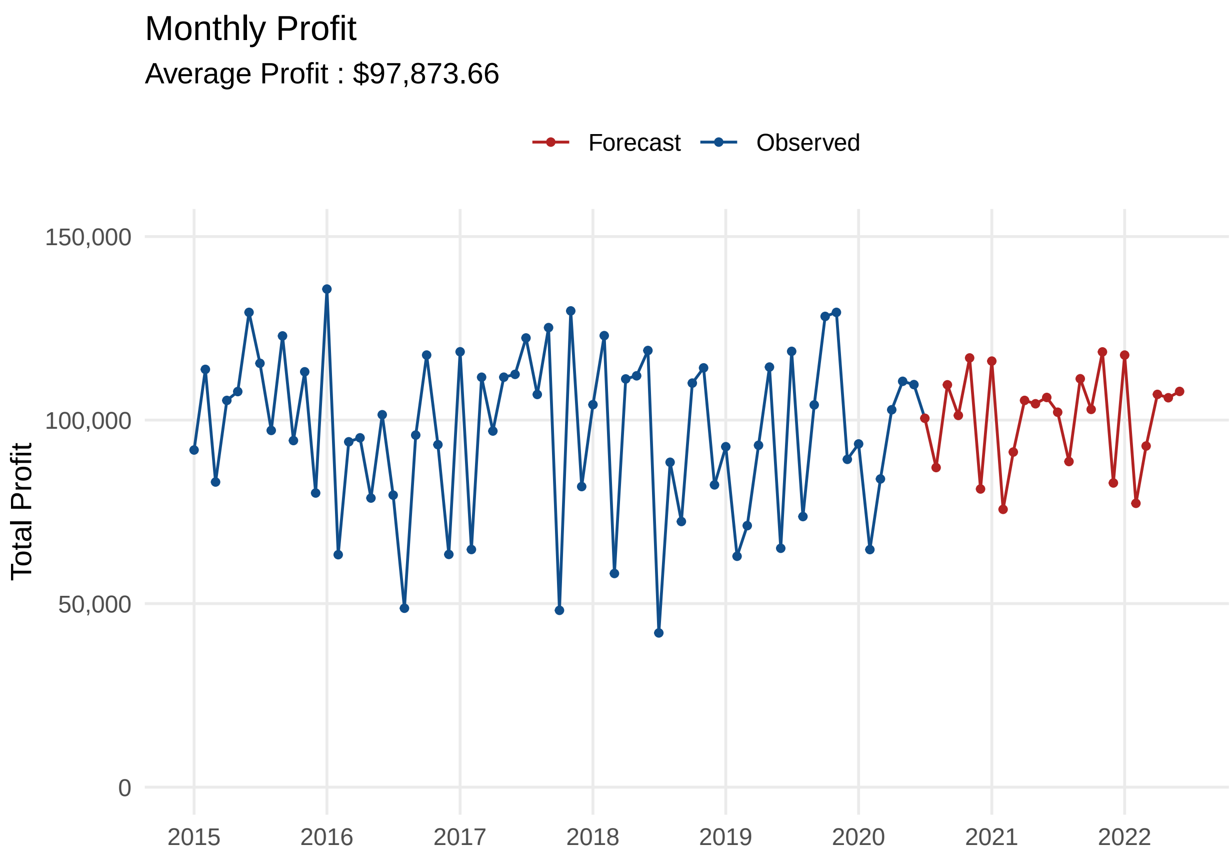
We will see how CRISP-DM applied in a business case study of increasing product sales by building a model to predict potential customers.



Business Understanding

Background

Our company sold an automotive product for over 20 years. However, for the last 5 years the monthly average profit has been constant and did not gain any significant growth. The condition will remain the same in the future if we do not do something. However, with limited member of sales team, we don't have enough resource to attract more customer. We want to be efficient instead of keep expanding the team, so we need another approach. With limited time and resources, we need to be able to quickly inspect and prioritize which customer is a potential buyer. We will also need to formally research on what makes them buy our products. By doing this, we can achieve higher or the same amount of profit with cheaper cost.



Business Goals and KPI

The business goal is determined together with other department. This part should be the continuation of the background problem.

- Gain insight on what drives people to buy our product
- Increase sales by 10% year over year
- Reduce annual marketing cost by 10%

Data Mining Goals and KPI

The data mining goal is determined by the data mining team and is a translation from the business goals.

- Build predictive model with 75% accuracy
- Build predictive model with 75% recall
- Build predictive model with 75% precision

Data Understanding

On the **Data Understanding** phase, we will gather, describe and explore the data to make sure it fits the business goal.

The deliverable or result of this phase should include:

- Data description
- Early data exploration report
- Data quality report

Gathering and Describing Data

Data are collected from the sales department in tabular format. The data consists of the past sales team interaction with the lead customer. The sales team keep record on whether the leads turn into purchase or refuse to buy the product, complete with the customer demographic information.

Here are some samples of the data.

```
##    flag gender       education house_val    age online customer_psy marriage
## 1     Y      M         4. Grad    756460  1_Unk      N            B
## 2     N      F         3. Bach    213171  7_>65      N            E
## 3     N      M 2. Some College    111147 2_<=25      Y            C
## 4     Y      M 2. Some College    354151 2_<=25      Y            B   Single
## 5     Y      F 2. Some College    117087  1_Unk      Y            J  Married
## 6     Y      F         3. Bach    248694 6_<=65      Y            B  Married
## 7     Y      M         3. Bach   2000000  1_Unk      Y            A  Married
## 8     N      F         3. Bach    416925 5_<=55      Y            C  Married
## 9     N      F           1. HS    207676 4_<=45      Y            G
## 10    Y      M           1. HS    241380  1_Unk      Y            C  Married
##    child    occupation mortgage house_owner    region fam_income
## 1      U  Professional     1Low                Midwest          L
## 2      U  Professional     1Low       Owner Northeast          G
## 3      Y  Professional     1Low       Owner   Midwest          J
## 4      U Sales/Service     1Low                   West          L
## 5      Y Sales/Service     1Low                  South          H
## 6      N  Professional     2Med       Owner      West          G
## 7      U  Professional     1Low              Northeast          C
## 8      Y  Professional     1Low       Owner     South          I
## 9      Y   Blue Collar     1Low      Renter      West          D
## 10     U Sales/Service     1Low              Northeast          G
```

We can also see try to get the information about the structure of the data including:

- The number of rows (each row represent a single customer data)
- The number of column
- The name of each column
- The data type of each column

```
## Rows: 40,000
## Columns: 14
## $ flag         <chr> "Y", "N", "N", "Y", "Y", "Y", "Y", "N", "N", "Y", "N",...
## $ gender       <chr> "M", "F", "M", "M", "F", "F", "M", "F", "F", "M", "M",...
## $ education    <chr> "4. Grad", "3. Bach", "2. Some College", "2. Some Coll...
## $ house_val    <int> 756460, 213171, 111147, 354151, 117087, 248694, 200000...
## $ age          <chr> "1_Unk", "7_>65", "2_<=25", "2_<=25", "1_Unk", "6_<=65...
## $ online       <chr> "N", "N", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y",...
## $ customer_psy <chr> "B", "E", "C", "B", "J", "B", "A", "C", "G", "C", "C",...
## $ marriage     <chr> "", "", "", "Single", "Married", "Married", "Married",...
## $ child        <chr> "U", "U", "Y", "U", "Y", "N", "U", "Y", "Y", "U", "Y",...
```

```
## $ occupation   <chr> "Professional", "Professional", "Professional", "Sales...
## $ mortgage     <chr> "1Low", "1Low", "1Low", "1Low", "1Low", "2Med", "1Low"...
## $ house_owner  <chr> "", "Owner", "Owner", "", "", "Owner", "", "Owner", "R...
## $ region       <chr> "Midwest", "Northeast", "Midwest", "West", "South", "W...
## $ fam_income   <chr> "L", "G", "J", "L", "H", "G", "C", "I", "D", "G", "E",...
```

The collected data consists of 40,000 distinct customers with 14 variables. The description of each column/variable can be seen below:

- **flag** : Whether the customer has bought the target product or not
- **gender** : Gender of the customer
- **education** : Education background of customer
- **house_val** : Value of the residence the customer lives in
- **age** : Age of the customer by group
- **online** : Whether the customer had online shopping experience or not
- **customer_psy** : Variable describing consumer psychology based on the area of residence
- **marriage** : Marriage status of the customer
- **children** : Whether the customer has children or not
- **occupation** : Career information of the customer
- **mortgage** : Housing Loan Information of customers
- **house_own** : Whether the customer owns a house or not
- **region** : Information on the area in which the customer are located
- **fam_income** : Family income Information of the customer(A means the lowest, and L means the highest)

Early Data Exporation and Data Quality Check

We also need to check the quality of the data. For example, since many of the column/variable is categorical, we can check the summary of the data and see the number of customer of each categories. By doing this, we can also check whether there are any data that need to be cleansed or to be transformed. For example, we can check if there is a missing/empty values.

The text above each section is the name of the column in the data. The text on the left side is the category on each column while the number on the right side is the frequency of each category. Numerical variable will be presented in summary statistics (mean, median, min, max, etc.).

```
##  flag      gender             education       house_val          age
##  N:20000   F:16830                    :  741   Min.   :      0   1_Unk :6709
##  Y:20000   M:22019   0. <HS          : 3848   1st Qu.:  80657   2_<=25:2360
##            U: 1151   1. HS           : 8828   Median : 214872   3_<=35:4984
##                      2. Some College:11400   Mean   : 307214   4_<=45:7115
##                      3. Bach         : 9267   3rd Qu.: 393762   5_<=55:8103
##                      4. Grad         : 5916   Max.   :9999999   6_<=65:5907
##                                                                  7_>65 :4822
##  online     customer_psy     marriage      child           occupation
##  N:12681   B      :8197          :14027   0:  127   Blue Collar  : 6621
##  Y:27319   C      :7830   Married:20891   N:13333   Farm         :  329
##            E      :6650   Single : 5082   U: 8528   Others       : 2006
##            F      :4058                   Y:18012   Professional :14936
##            G      :3951                             Retired      : 4341
##            D      :2353                             Sales/Service:11767
##            (Other):6961
##   mortgage     house_owner          region        fam_income
```

```
##  1Low :29848         : 3377   Midwest  : 8107   E      : 8432
##  2Med : 4803   Owner :29232   Northeast: 7247   F      : 6641
##  3High: 5349   Renter: 7391   Rest     :  245   D      : 4582
##                               South    :15676   G      : 4224
##                               West     : 8725   C      : 2687
##                                                 H      : 2498
##                                                 (Other):10936
```

We can check the full summary for `customer_psy` and `fam_income` column since they contain many categories.

```
## Customer Psychology
##    A    B    C    D    E    F    G    H    I    J    U
## 1427 8197 7830 2353 6650 4058 3951  958 2262 2187  127
```

```
## Family Income
##    A    B    C    D    E    F    G    H    I    J    K    L    U
## 2274 2169 2687 4582 8432 6641 4224 2498 1622 1614 1487 1617  153
```

There are some interesting finding from the summary. For example, the `gender` column consists of 3 categories: F (Female), M (Male), and U (Unknown). The `child` column is similar, with additional value of U (Unknown) and 0 (zero) even though the column should only be Yes or No. The `marriage` and `education` column contain empty values. This is not surprising, since the sales team are not instructed to fulfill each column with pre-determined values. However, this means that the incoming data quality is not good and require future standardization in the future. This also show us that we need to cleanse and prepare the data before we do any analysis so that all relevant information can be captured.

Data Preparation

On the **Data Understanding** phase, we will prepare and cleanse the data so they are fit for analysis and making prediction. Some people said that the data preparation take 80% of the data mining process.

The deliverable or result of this phase should include:

- Data preparation steps
- Final data for modeling

Data Cleansing

On this process, we handle the data based on the problem we find during the data understanding phase. Based on our finding, we will do the following process:

- Change missing/empty value in `education`, `house_owner` and `marriage` into explicit Unknown
- Make all *U* value in all categorical column into explicit Unknown
- Cleanse the `age` category by removing the index (1_Unkn into Unknown, 2_<=25 into <=25, etc.)
- Cleanse the `mortgage` category by removing the index

```
##   flag           gender             education        house_val
##  No :20000   Female :16830   <HS         : 3848   Min.   :      0
##  Yes:20000   Male   :22019   Bach        : 9267   1st Qu.:  80657
```

```
##              Unknown: 1151   Grad        : 5916   Median : 214872
##                              HS          : 8828   Mean   : 307214
##                              Some College:11400   3rd Qu.: 393762
##                              Unknown     :  741   Max.   :9999999
##
##       age        online       customer_psy     marriage         child
##  <=25   :2360   No :12681   B      :8197   Married:20891   No     :13333
##  >65    :4822   Yes:27319   C      :7830   Single : 5082   Unknown: 8655
##  26-35  :4984               E      :6650   Unknown:14027   Yes    :18012
##  36-45  :7115               F      :4058
##  46-55  :8103               G      :3951
##  56-65  :5907               D      :2353
##  Unknown:6709               (Other):6961
##         occupation    mortgage      house_owner          region
##  Blue Collar  : 6621   High: 5349   Owner  :29232   Midwest  : 8107
##  Farm         :  329   Low :29848   Renter : 7391   Northeast: 7247
##  Others       : 2006   Med : 4803   Unknown: 3377   Rest     :  245
##  Professional :14936                                South    :15676
##  Retired      : 4341                                West     : 8725
##  Sales/Service:11767
##
##    fam_income
##  E      : 8432
##  F      : 6641
##  D      : 4582
##  G      : 4224
##  C      : 2687
##  H      : 2498
##  (Other):10936
```

Now that the data is already cleansed, we need to consider whether we need to remove data that contain any *Unknown* value? Should the sales team need to know all information about a customer to make a prediction or are they allowed to fill some variable with Unknown? In this step we need to discuss with the sales team since they are the final user of the model.

Let's say together with the sales team we have decided that any data that contain missing value should not be used for analysis Therefore, we will drop/remove any row/customer that has missing information about them.

Finally, after careful and rigorous data cleansing, we acquire our final data that will be used for analysis and modeling.

Final Data

```
##    flag gender    education house_val   age online customer_psy marriage child
## 1   Yes Female         Bach    248694 56-65    Yes            B  Married    No
## 2    No Female         Bach    416925 46-55    Yes            C  Married   Yes
## 3   Yes Female Some College    360587 46-55    Yes            C  Married   Yes
## 4    No Female           HS         0   >65     No            I  Married    No
## 5   Yes Female Some College    239560 46-55    Yes            C  Married   Yes
## 6    No Female Some College    136729 36-45    Yes            C  Married   Yes
## 7   Yes   Male         Bach    308817 26-35    Yes            C  Married   Yes
## 8   Yes   Male         Grad    206271 26-35    Yes            C  Married   Yes
## 9   Yes   Male Some College    169113 56-65    Yes            B  Married    No
```

```
## 10  Yes   Male           HS   1076582 46-55    Yes           B  Married   Yes
##       occupation mortgage house_owner  region fam_income
## 1   Professional      Med       Owner    West          G
## 2   Professional      Low       Owner   South          I
## 3   Professional     High       Owner Midwest          J
## 4        Retired      Low       Owner   South          E
## 5  Sales/Service      Med       Owner Midwest          F
## 6    Blue Collar      Low       Owner Midwest          G
## 7  Sales/Service      Low      Renter   South          F
## 8   Professional      Med       Owner    West          F
## 9   Professional      Low       Owner Midwest          F
## 10  Professional     High       Owner    West          F
```

Data Understanding (Again)

As expected, CRISP-DM is not a linear process. We can go back and forth between process to make sure it fits the business and data mining goal. Here, we go back to data understanding phase to further explore and analyze the data before we start to make a machine learning model.

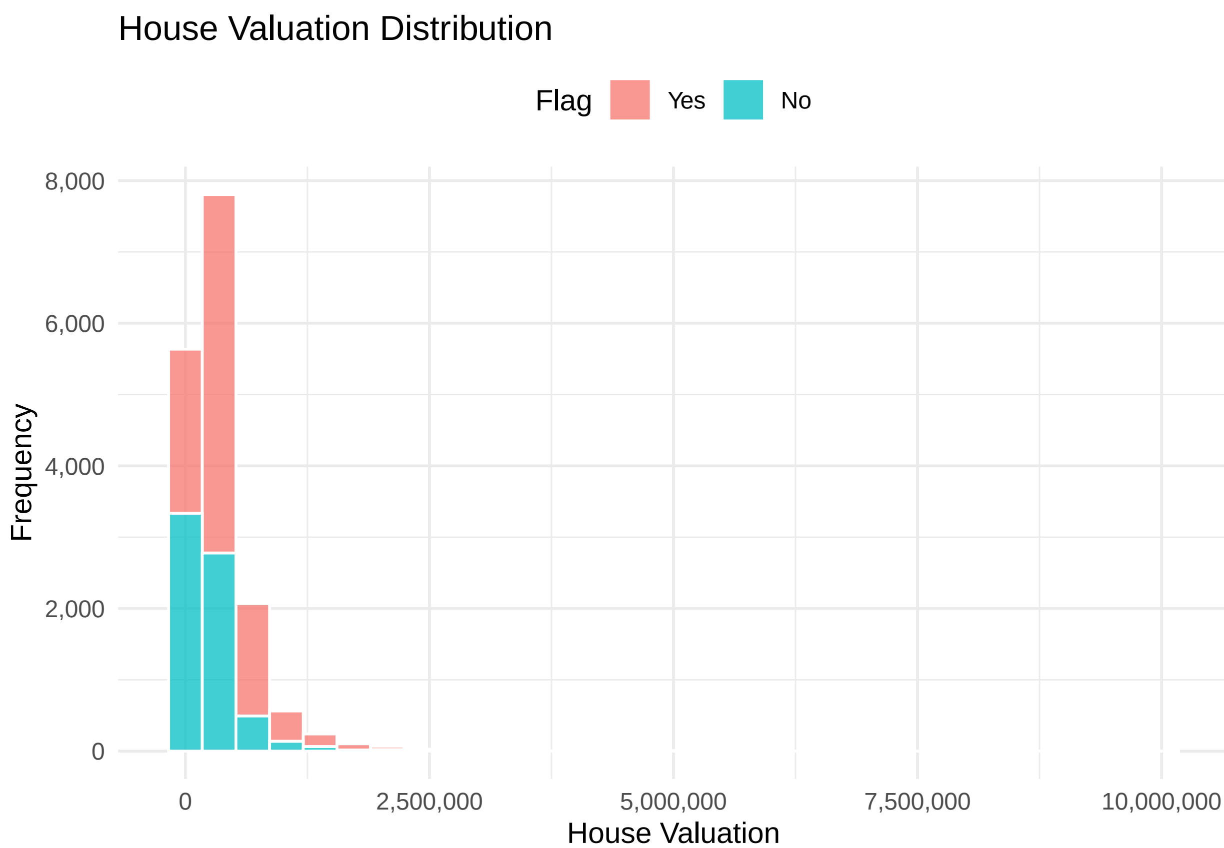
Exploratory Data Analysis

The process of exploring and visualizing insight from the data is called **Exploratory Data Analysis (EDA)**.

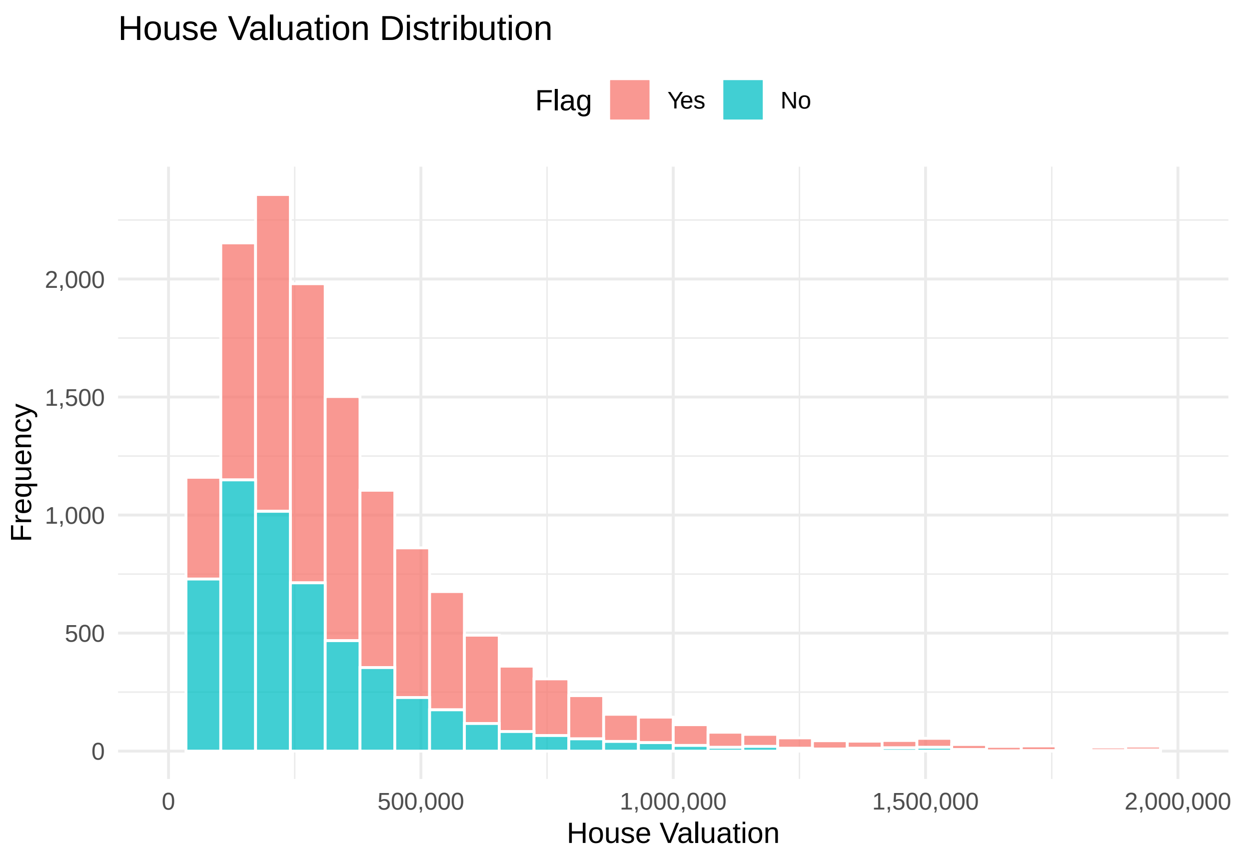
House Valuation Distribution

Here we will do visualization to see whether there are any difference between customer who buy our product and who don't. To visualize a distribution, we can use histogram. The *x-axis* is the house valuation while the *y-axis* show the frequency or the number of customer with certain house valuation.

From the histogram, most of our customer has house valuation less than 2,500,000. Some customers are outlier and has house valuation greater than 2,500,000. Their frequency is low and they cannot be seen on the histogram. The distribution for people who buy and not buy are quite similar, therefore we cannot simply decide if a customer will buy our product based on their house valuation.



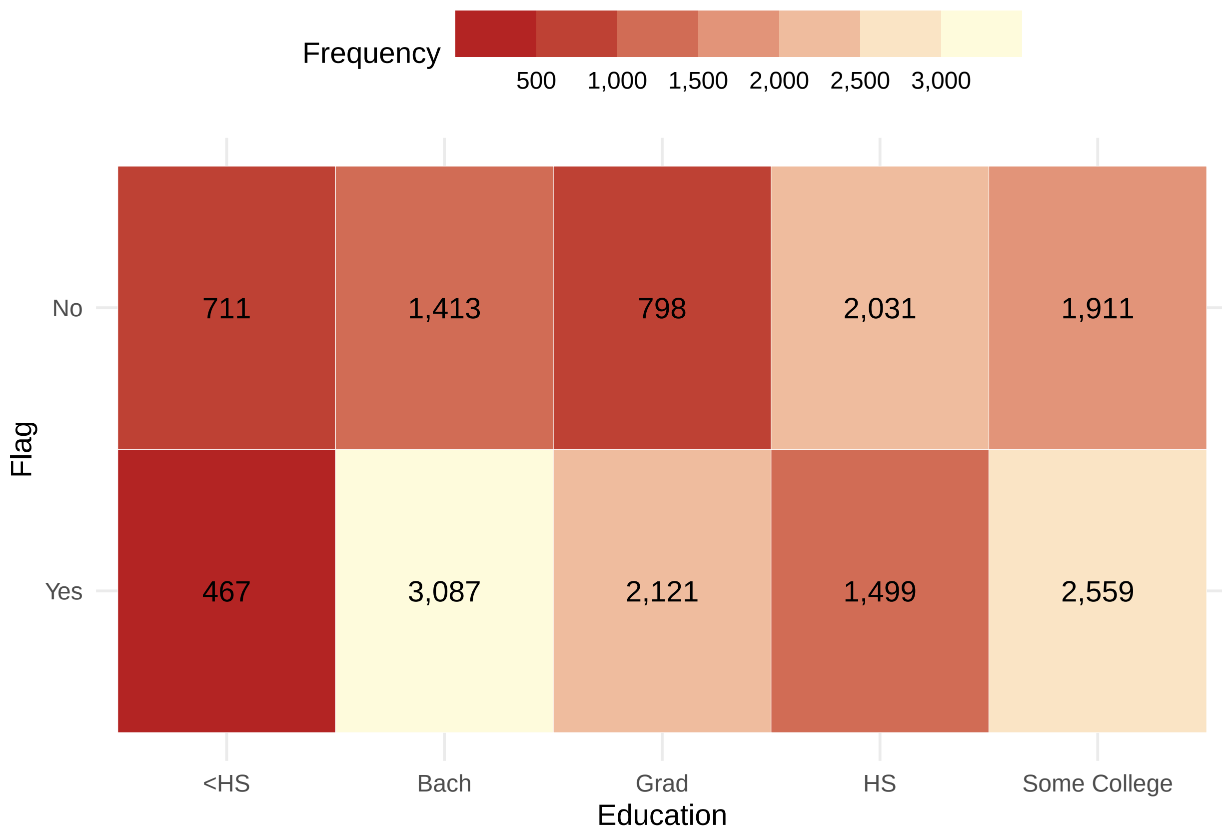
We can cut and remove the outlier to see the distribution better.



Education Level

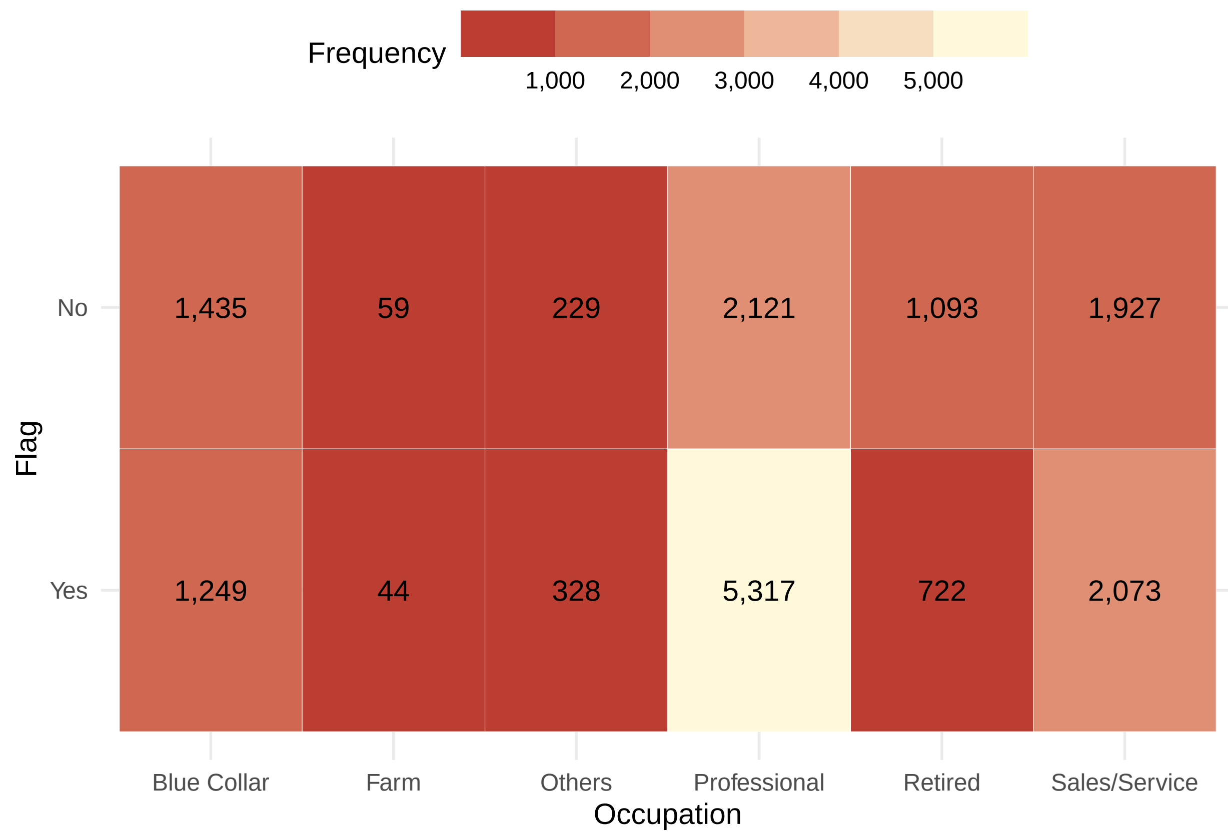
We will see if the education level can be a great indicator to decide if a customer has high probability to buy our product. The color of each block represent the frequency of people that fell in that category, with brighter color indicate higher frequency.

Based on the heatmap, people with higher education level (*Bach* and *Grad*) are more likely to buy our product. Therefore, education level may be a great indicator to check potential customer.



Occupation

We will do the same thing here with the occupation/job. The one that stands out is the professional occupation that has a very high frequency of people who buy our product.



You can keep doing exploratory with other variables and with different approach. The point of EDA is to make understand more about the data and finding new insight before making a predictive model.

Modeling

On the **Modeling** phase, we will start creating model to find pattern inside our data and to make future prediction for business purpose.

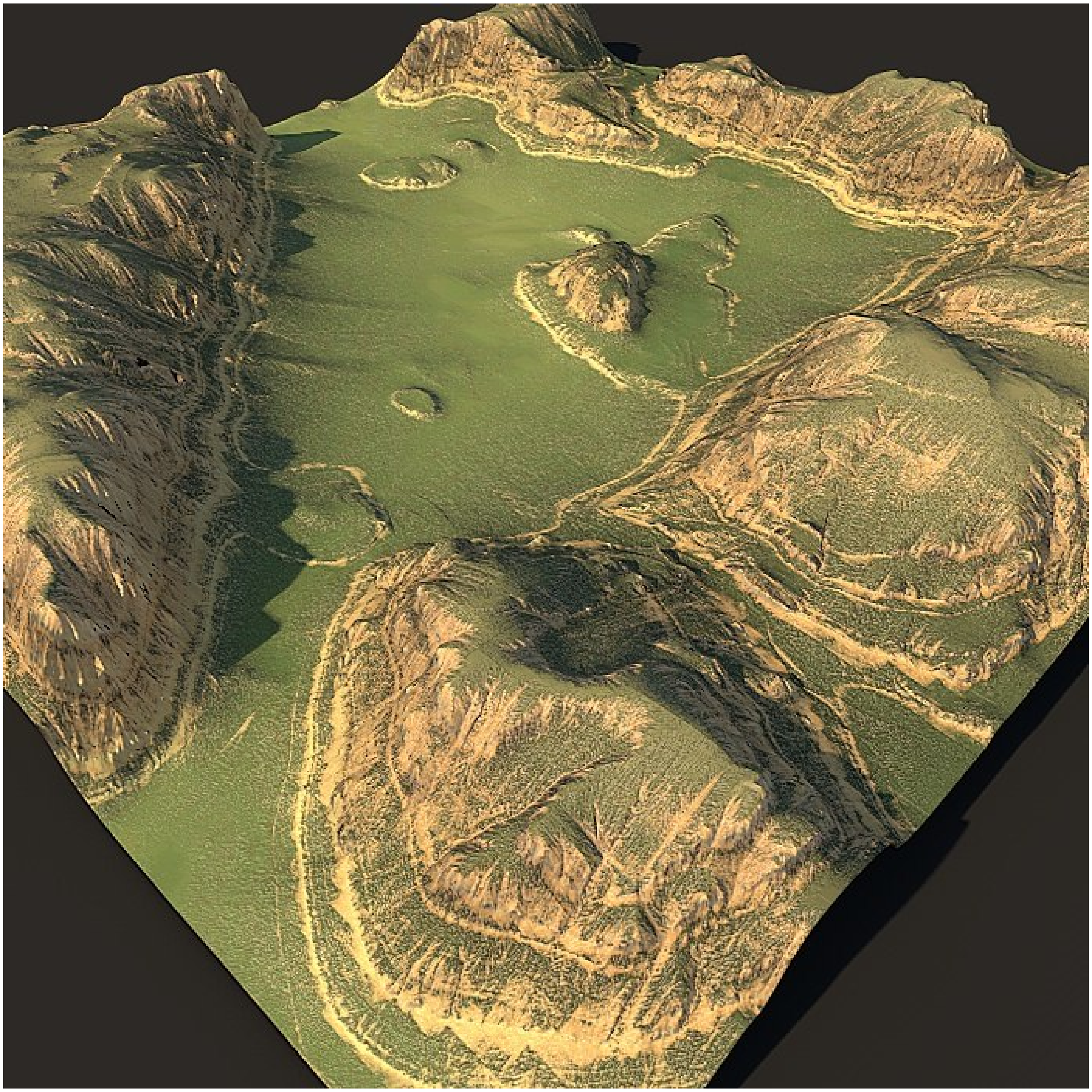
The deliverable or result of this phase should include:

- Modeling Technique and assumption
- Model Description
- Model Evaluation

Model and Machine Learning

A model is a representation of the world. Since it's just a representation, some information may lost and not very accurate. However, the model still useful for some purpose.

> All models are wrong, but some are useful. - George Box

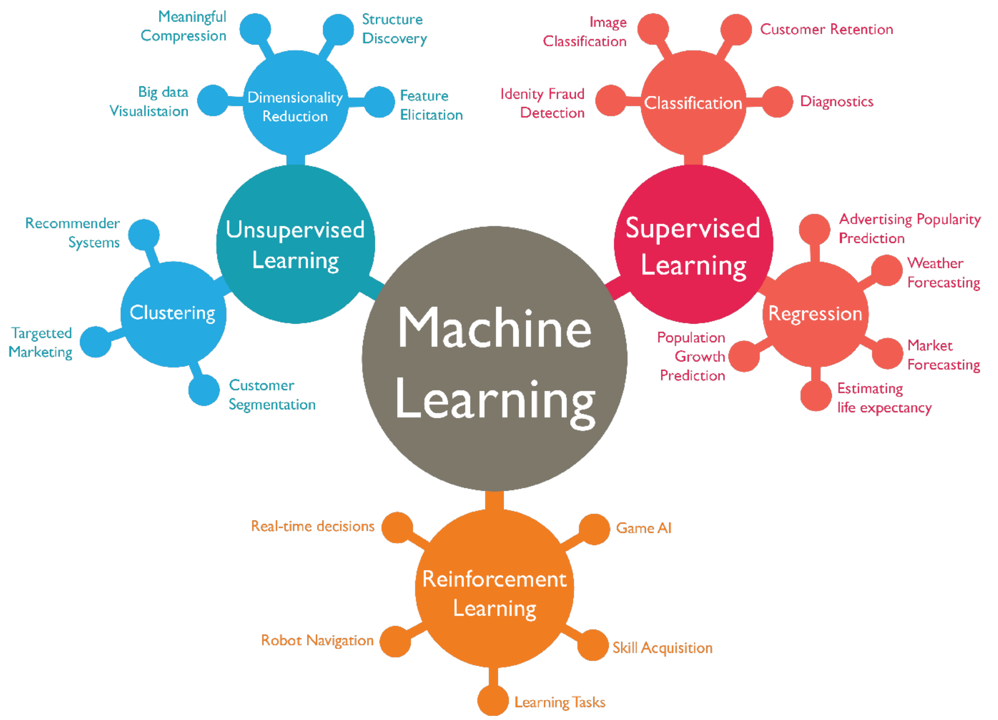


Machine learning is a statistical model that is specifically designed to learn and find pattern from a data. The data that is used to train the model is called the **Training Dataset**. Depending on their purpose, they can be divided into several categories:

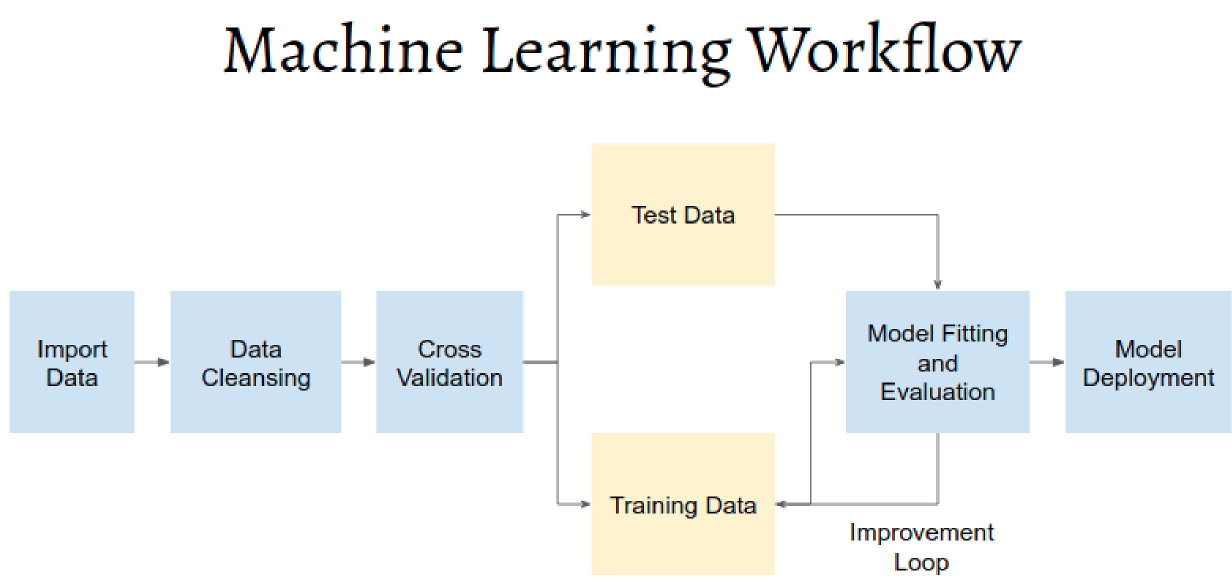
- **Supervised Learning**: Model learn and being supervised. There is a target variable, a variable we want to predict. Imagine the model as a student who learn from a data, then they need to make a prediction. If the model is wrong, it will try to correct itself until the error is minimum.
  - **Regression problem** is where we want to predict a numerical variable, such as a house price, car price, energy consumption, etc.
  - **Classification problem** is where we want to predict a categorical variable, such as the probability of a customer to churn, detecting cancer cell from image, credit scoring, etc.

- **Unsupervised Learning**: There is no target variable. Model is free to find its own pattern.
- **Reinforcement Learning**: Model learn by interacting with the environment. The model often used for simulation and decision making.

Below is some application of each respective category of machine learning, with no specific machine learning algorithm being mentioned.



Machine learning is part of the data mining process. However, we will illustrate the general machine learning workflow with this simple figure.



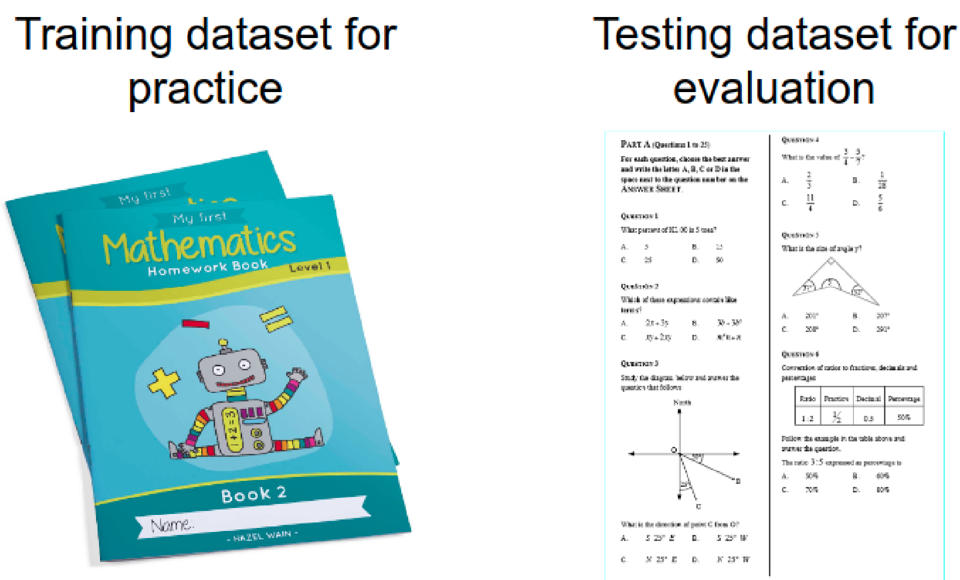
The import data and the data cleansing is the same process as before. The next step for building a model is to do a process called Cross-Validation.

Cross-Validation

The cross-validation step is where we will split our data into 2 separate dataset: training dataset and testing dataset.

- **Training Dataset**: Dataset that will be used to train the machine learning model
- **Testing Dataset**: Dataset that will be used to evaluate the performance of the model.

Why do we need to separate the data? Because the model will always perform better in the data that they've trained with. Imagine where you are doing a math homework. You can easily do them, especially after you check the correct answer and learn what makes you wrong. However, we want our model to be able to predict a new, unseen data. That's why we need the testing dataset. The testing dataset acts as the examination or evaluation for the model, to check whether they can truly learn the pattern inside the data.



Here we split the data with 80% of the data will be the training dataset and the rest will be the testing dataset. Each observation/row is randomly selected as either the training set or the testing set. The random selection is done to make sure we don't include any selection bias done by human.

The data train consists of 13,279 rows while the data test consists only of 3,318 rows.

```
## Number of Data Train :  13279
```

```
## Number of Data Test :  3318
```

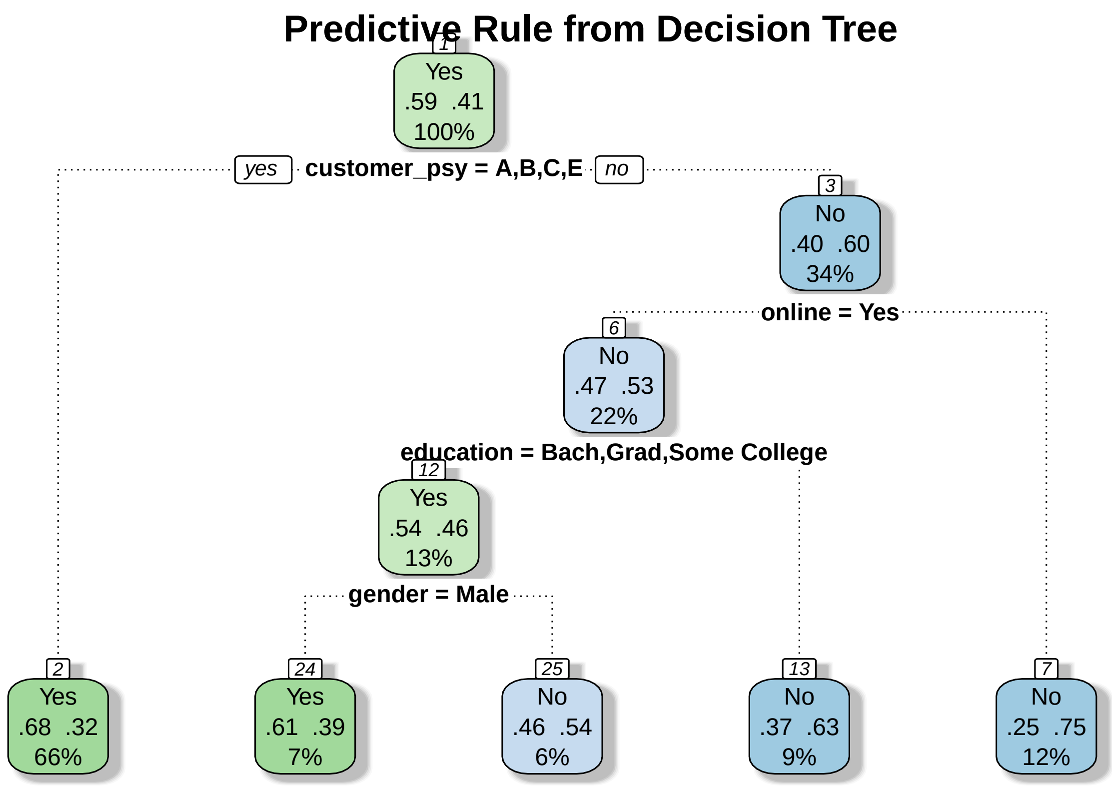
Model Fitting

Here, we fit or train the model using the data train. We will use 2 different models: Decision Tree and Random Forest. Later, we will evaluate both models and choose only the best model.

Decision Tree

The decision tree is a machine learning algorithm that try to create a set of rule to classify and predict the target variable. The model tries to split the data into homogeneous group based on the predictor variable.

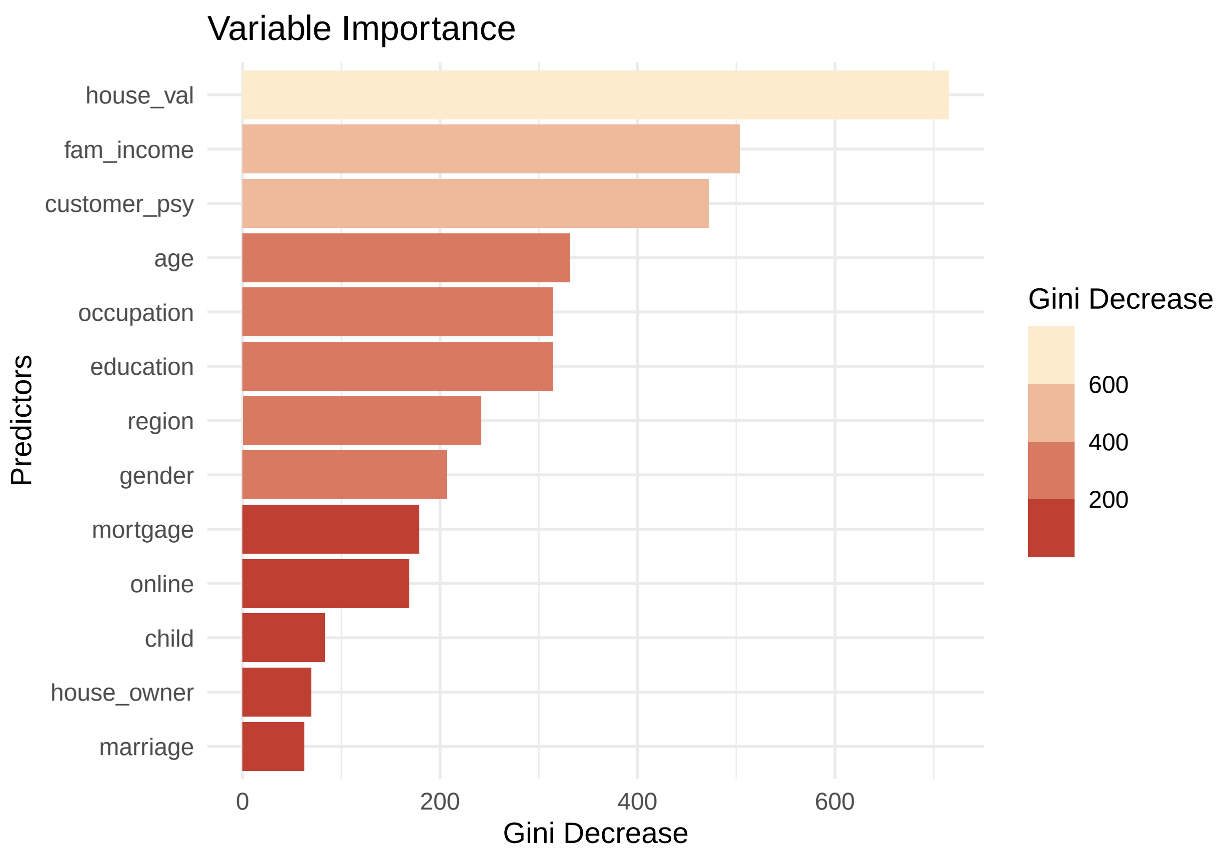
You can see from the figure below that decision tree is like a flow chart that we can actually follow. For example, if the customer psychology is either A, B, C, or E, then there is a high chance that the customer will buy our product. If the customer don't belong in those customer psychology, we then proceed to check whether the customer had any previous online experience. If he/she never had any online experience, then it is more likely that the customer will not buy our product. The higher variable has higher importance in determining customer's buying decision. By understanding which factors are important, we can provide better service and promotion for customer to increase the chance of conversion.



Random Forest

The next model is Random Forest. In short, Random Forest is a collection of decision tree that together make a single decision. Imagine you are in a middle of a presidential election and as a country you need to decide which presidential candidate to choose. Each citizen is a single decision tree with their own prediction. Together, they decide which candidate that will be elected and the final decision is the majority voting. Random Forest is more powerful than Decision Tree due to this characteristics.

However, we can't get a nice plot of flowchart like the previous decision tree. Instead, we can only get the importance of each variable based on a certain metric called *Gini Index*. According to the Random Forest, the most important variable to predict customer's buying decision is the `house valuation`, followed by the `family income` and `customer psychology`.



Now that we understand and describe the model, we need to evaluate them and see if they are actually able to distinguish customer that buy our product.

Model Evaluation

In classification problem, we evaluate model by looking at how many of their predictions are correct. This can be plotted into something called **Confusion Matrix**.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

The matrix is divided into four area:

- **True Positive (TP)**: The model predict customer **will buy** and the prediction is correct (customer buy)
- **False Positive (FP)**: The model predict customer **will buy** and the prediction is incorrect (customer not buy)
- **True Negative (TN)**: The model predict customer **will not buy** and the prediction is correct (customer not buy)
- **False Negative (FN)**: The model predict customer **will not buy** and the prediction is incorrect (customer buy)

For example, here is the confusion matrix from the decision tree after doing prediction to the testing dataset.

```
##           Truth
## Prediction  Yes   No
##        Yes 1648  791
##        No   298  581
```

If we define positive as Yes:

- **True Positive (TP)**: 1648
- **False Positive (FP)**: 791
- **True Negative (TN)**: 581
- **False Negative (FN)**: 298

Next, we can start doing evaluation using 3 different metrics: accuracy, recall, and precision. Those metrics are pretty general and complement each other. There are more evaluation metrics but we will not be discussed it here.

- **Accuracy**

Accuracy simply tell us how many prediction is true compared to the total dataset.

Accuracy

Use all value

	Actual Yes	Actual No
Predicted Yes	TP	FP
Predicted No	FN	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

```
(1648 + 581) / (1648 + 581 + 791 + 298)
```

```
## [1] 0.6717902
```

From all data in testing dataset, only 67% of them are correctly predicted as buy/not buy.

$$Accuracy = \frac{1648 + 581}{1648 + 581 + 791 + 298} = 0.67179 = 67.18\%$$

- **Sensitivity/Recall**

Recall/sensitivity only concerns how many customers that actually buy can correctly be predicted. The metric don't care about the customer that don't actually buy our product.

Recall

Only use value with Actual Yes

	Actual Yes	Actual No
Predicted Yes	TP	FP
Predicted No	FN	TN

$$Recall = \frac{TP}{TP + FN}$$

```r
1648 / (1648 + 298)
```

```
## [1] 0.8468654
```

From all customer that actually buy our product, 84% of them are correctly predicted as buy and 16% as not buy.

$$Recall = \frac{1648}{1648 + 298} = 0.8468 = 84.68\%$$

- **Precision**

Precision only concern on how many positive prediction that are actually correct. The metric don't care about customer that is predicted not buy.

Precision

Only use value with Predicted Yes

	Actual Yes	Actual No
Predicted Yes	TP	FP
Predicted No	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

```r
1648 / (1648 + 791)
```

```
## [1] 0.6756868
```

From all customer that is predicted to buy, only 67% of them that are actually buy our product.

$$Precision = \frac{1648}{1648 + 791} = 0.67568 = 67.57\%$$

Decision Tree

Here is the recap of the evaluation metrics for Decision Tree.

```
##    Accuracy    Recall Precision
## 1 0.6717902 0.8468654 0.6756868
```

Random Forest

Here is the recap of the evaluation metrics for Random Forest. The model is slightly better than the Decision Tree.

```
##    Accuracy    Recall Precision
## 1 0.6865582 0.8016444  0.704607
```

With that, we can go back to the business goal, specifically the **Data Mining Goals** of our project. Does our model have achieved our data mining goals?

- Build predictive model with 75% accuracy
- Build predictive model with 75% recall
- Build predictive model with 75% precision

If the model doesn't satisfy the data mining goals, the team can works on improving the model to get better performance. That's why we have an improvement loop on the machine learning workflow. We rarely achieve our best model on the first run and need to do several iterations on improvement until we find the best model.

For now, we will proceed to the next step.

Evaluation

On the **Evaluation** phase, we will further evaluate the model into the context of the business problem.

The deliverable or result of this phase should include:

- Model business assesment
- Review of the overal process
- Possible action and final decision

Cost and Benefit Analysis

The cost and benefit analysis is where we try to convert the machine learning performance into the business context. We will try to see by employing the machine model, how many profit that we can make compared to the average profit we currently have?

Define Cost and Benefit

The first we do is to define the cost and benefit of each decision. We will define it similar with the previous confusion matrix. The main cost is the cost of approaching a customer, in here we defined it as 600. The revenue generated for each customer is 1000, with the profit of 400 after we cut the revenue with the cost.

- **True Positive (TP)**: If the model predict customer **will buy** and the prediction is correct (customer buy), we will get a profit of 400 (1000 revenue - 600 cost)
- **False Positive (FP)**: If the model predict customer **will buy** and the prediction is incorrect (customer not buy), we will lost 600
- **True Negative (TN)**: If the model predict customer **will not buy** and the prediction is correct (customer not buy), nothing happened
- **False Negative (FN)**: If the model predict customer **will not buy** and the prediction is incorrect (customer buy), nothing happened

```
##   Benefit True Positive Benefit True Negative Cost False Positive
## 1                   400                     0                -600
##   Cost False Negative
## 1                   0
```

Profit Curve

We will prioritize the customer that has the highest probability to buy our product. Thus, first we make a list of a high scoring customer.

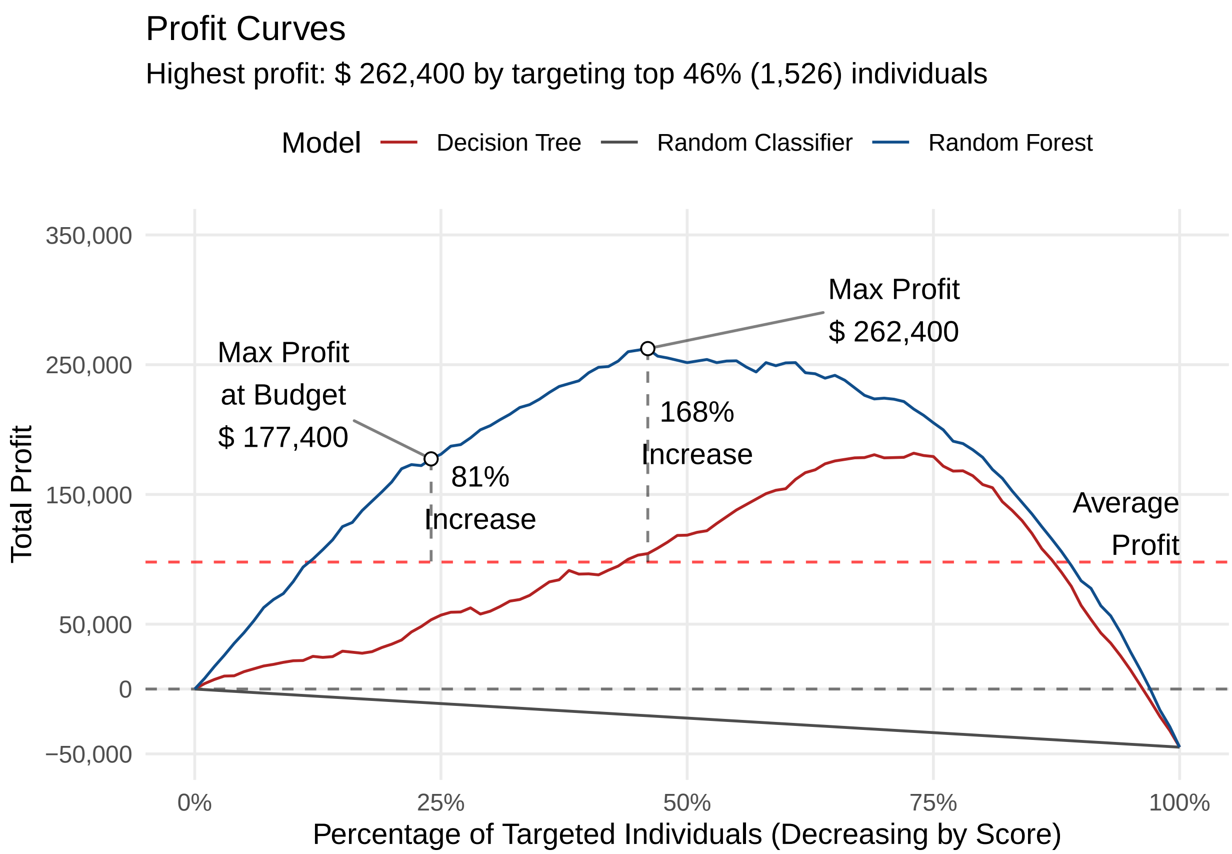
```
##      Yes    No prediction truth
## 1  1.000 0.000        Yes   Yes
## 2  1.000 0.000        Yes   Yes
## 3  1.000 0.000        Yes    No
## 4  1.000 0.000        Yes   Yes
## 5  1.000 0.000        Yes   Yes
## 6  1.000 0.000        Yes   Yes
## 7  0.998 0.002        Yes   Yes
## 8  0.998 0.002        Yes   Yes
## 9  0.998 0.002        Yes    No
## 10 0.998 0.002        Yes   Yes
```

And we calculate for how many profit we will get if we target and approach only some percent of the total customer? For example, if we only target the top 10 customer and ignore the rest, we have correctly predict 9 customer as buy and only a single incorrect prediction (the actual buying decision is on the `truth` column). Thus, our profit would be:

$$Profit = 9 \times 400 + (-600) = 3000$$

Since we have 0 cost and 0 benefit for negative prediction, we can skip the calculation and our final profit is only a mere 3000.

That's how we will calculate the profit. We will do the same thing but calculate some level of percent of people that will be targeted. The final result is the following profit curve, which shows the total profit that can be generated by targeting the top % of the customer.

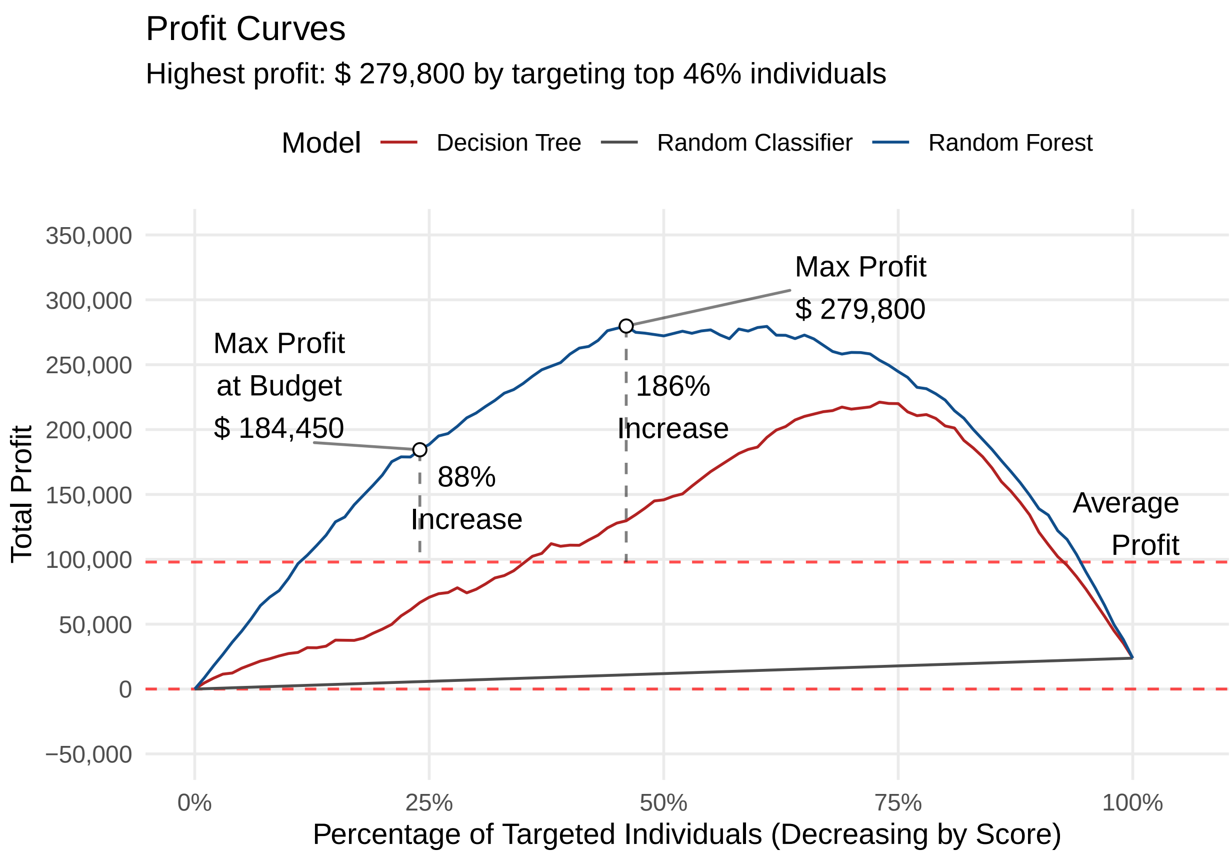


The maximum profit that we can get is 262,400 by targeting the top 46% individuals using the score from Random Forest. Compared to the average profit we gain every month, this is a 186% increase. We also show you how many profit generated by random classifier, which is just a simple random guess (50:50 probability).

We can also add some scenarios. For example, the sales has only a monthly budget of 480,000. If each customer cost 600 to approach, we can only target 480,000/600 = 800 individuals = top 24% leads. In this scenario, we will get 177,400, which is still a big improvement from our average profit. The main focus of this profit curve is to show you that we don't need to approach all customer and prioritize the one that has the highest chance to buy our product.

If we change the cost and benefit value, the graph will also change. For example, let's say we have successfully cut the marketing cost by 50 from 600 to 550.

```
##   Benefit True Positive Benefit True Negative Cost False Positive
## 1                   400                     0                -550
##   Cost False Negative
## 1                   0
```



Review Overal Process

The overal process of the data mining is quite smooth with some flaws that we find:

- Dirty or improper input data
- Underperforming model
- Data gathering is not done in real time yet

Final Decision

After reviewing the project, there are some possible action for us to do:

- Improve the model before release them into the real use
- Release the model while also developing a better model
- Create a standardized data input procedure
- Present a full report of the data mining project

Deployment

Deployment is where data mining pays off. It doesn't matter how brilliant your discoveries may be, or how perfectly your models fit the data, if you don't actually use those things to improve the way that you do business. We have build the model but how do we use them in real life situation? We can launch the model in several ways. The common method to release a machine learning model into production/real use case is as follows:

- Building a dashboard
- Building an API

The deliverable or result of this phase should include:

- Deployment plan
- Monitoring and Maintenance
- Final Report

Conclusion

The CRISP-DM methodology is a long and detailed standard that will make your data mining project fits your business needs and documented properly. Its not a linear method, you can always go back to the previous step if you found an issue or need to renew some goals and process.