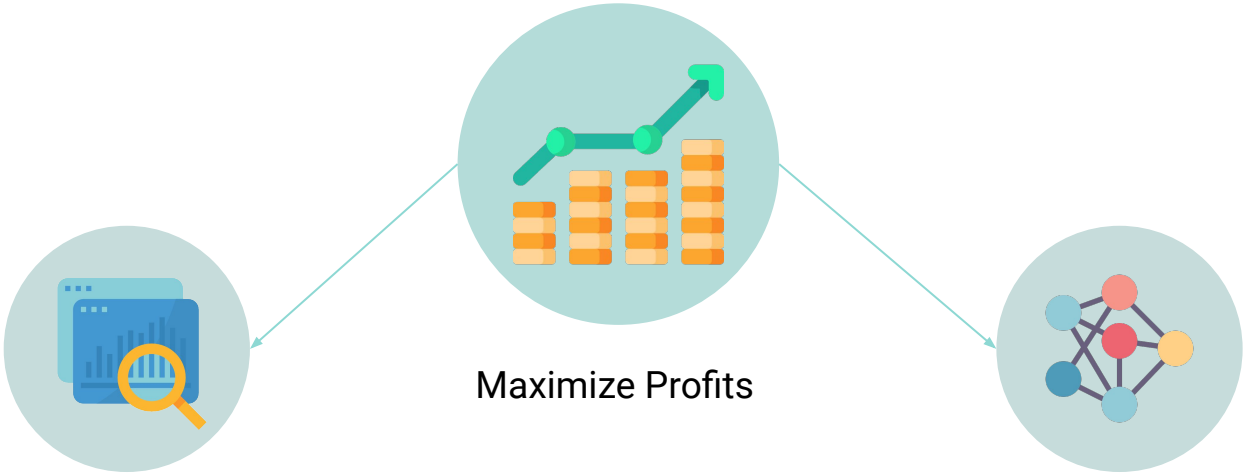
Marketing Campaign Analysis & Prediction

IDS 400 Final Project

Created by:

- Anh Pham
- Arifa Baber
- Francisco Sanchez
- Jing Wen Chui
- Joseline Tanujaya

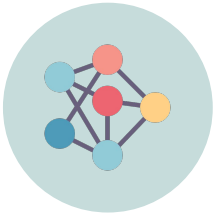
Objective


Maximize Profits
Understand your customers
Predictive Models

Business Questions

Visualization

- Define the profile of a potential customers? (Who are more likely to be interested in our business)
- Observe the customers behavior for future strategic decision making



Machine Learning

- Create machine learning algorithms that can predict who are the customers who are more likely to response the next marketing campaign

About our data

Source: https://www.kaggle.com/rodsaldanha/arketing-campaign

Contains 2,240 observations of 29 initial variables

Records the customer’s demographic information (excluding geographical), their purchasing behavior, and their history of responses to previous campaigns.

A glimpse of data

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
0	5524	1957	Graduation	Single	58138.0	0	0	2012-09-04	58	635	88	546	172
1	2174	1954	Graduation	Single	46344.0	1	1	2014-03-08	38	11	1	6	2
2	4141	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	426	49	127	111
3	6182	1984	Graduation	Together	26646.0	1	0	2014-02-10	26	11	4	20	10
4	5324	1981	PhD	Married	58293.0	1	0	2014-01-19	94	173	43	118	46
5	7446	1967	Master	Together	62513.0	0	1	2013-09-09	16	520	42	98	0
6	965	1971	Graduation	Divorced	55635.0	0	1	2012-11-13	34	235	65	164	50
7	6177	1985	PhD	Married	33454.0	1	0	2013-05-08	32	76	10	56	3
8	4855	1974	PhD	Together	30351.0	1	0	2013-06-06	19	14	0	24	3
9	5899	1950	PhD	Together	5648.0	1	1	2014-03-13	68	28	0	6	1

Preparation

First step to the analysis

Preparation & Cleaning

- Drop ID column as it will cause leakage in modeling process

```
df['ID'].value_counts().index.sort_values(ascending=True)
df.drop('ID',axis = 1, inplace = True)
```

Derive New Variables

Customer's age as per data collection time

```
current_date = date.today()

df['Age']=current_date.year-df['Year_Birth']
```

Spending (Total spending across all categories)

```
df['Spending']=
df['MntWines']+df['MntFruits']+df['MntMeatProducts']+df['MntFishProducts']+df['MntSweetProducts']+df['MntGoldProds']
```

Binary variable: to identify if a household has a child/not

```
df['Has_child'] =
np.where(df.Kidhome+df.Teenhome > 0, 'Yes', 'No')
```

Create ordinal variable for Marital_Status (assign codes to categorize them)

1: Alone, Single, Widow, Yolo , Absurd
2: Divorced
3: Married
4. Together

Create ordinal variable for Education

1: 2nd Cycle
2: Graduation
3: PhD
4. Master
5. Basic

Sum the customer's responses to previous campaigns to analyse purchase behavior

```
df['Sum Response'] = df['AcceptedCmp1'] +
df['AcceptedCmp2'] + df['AcceptedCmp3'] +
df['AcceptedCmp4'] + df['AcceptedCmp5'] +
df['Response']

df['Response_5campaigns'] =
df['AcceptedCmp1'] + df['AcceptedCmp2'] +
df['AcceptedCmp3'] + df['AcceptedCmp4'] +
df['AcceptedCmp5']
```

Create variable of types of customers based on their responses:

Repeat/Just Converted/ Not Converted/ Not Responding

```
def custtype(x):
if x['Response_5campaigns'] > 0 and x['Response'] == 1:
    return 'Repeat'
  elif x['Response_5campaigns'] == 0 and x['Response'] == 1:
    return 'Just Converted'
  elif x['Sum Response'] == 0:
    return 'Not Converted'
  elif x['Response_5campaigns'] > 0 and x['Response'] == 0:
    return 'Not Responding'

df['Type'] = df.apply(custtype, axis = 1)
```

The age at which a customer joins the company list

The number of days a customer has been with the company

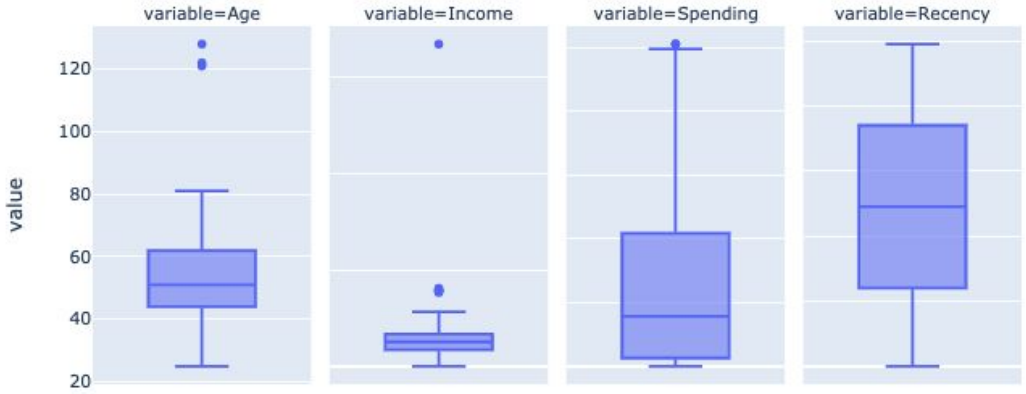
```
df['Dt_Customer'] =
pd.to_datetime(df['Dt_Customer'], format =
"%Y-%m-%d")
x = pd.to_datetime('2014-12-31', format =
"%Y-%m-%d")
for i in range(0, len(df)):
   df['Days_with_company'] = (x -
df['Dt_Customer']).dt.days
```

Impute Missing Data (KNN Method)

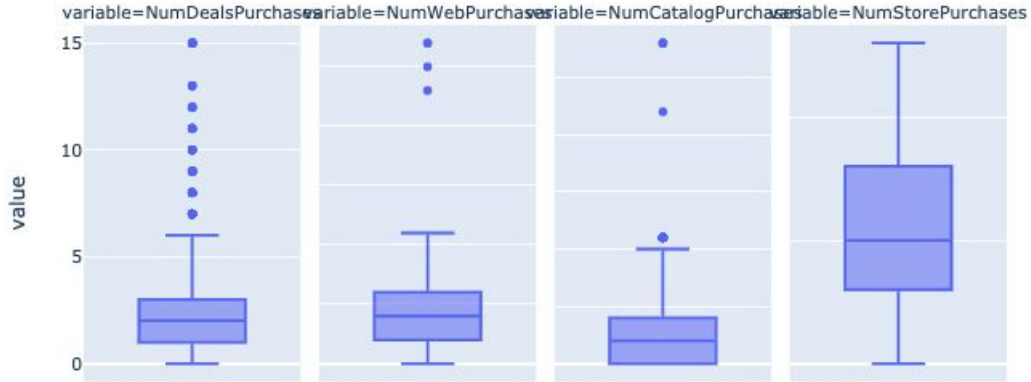
- KNNImputer (n_neighbors=5)

	Number of Missing Income Data	Proportion of Missing data against data set
Income	24	0.010714

Remove Outlier (Income Variable)



- Outlier observation found in the income variable: removed

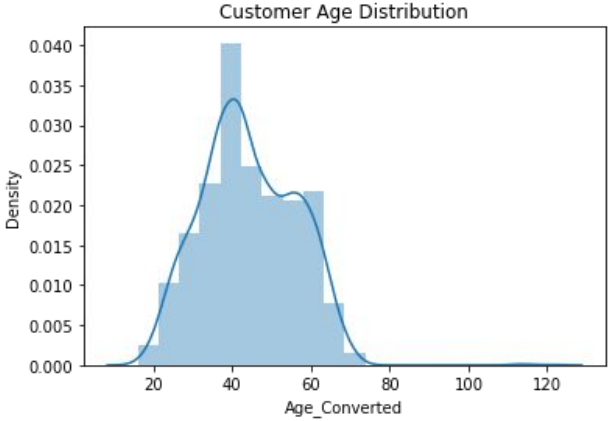


Understand The Business

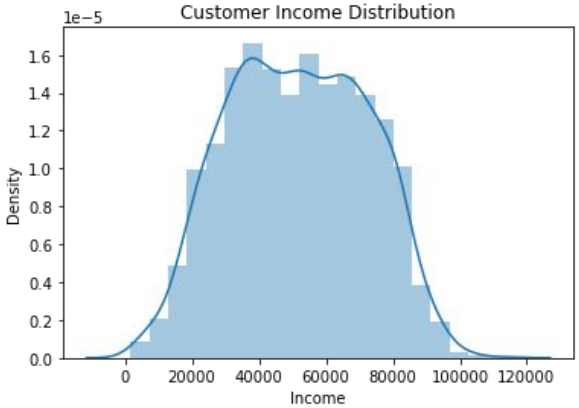
Through visualization to guide future business strategy

Exploratory Data Analysis

Of demographic variables



mean	44.235215
std	12.029053
min	16.000000
25%	36.000000
50%	43.000000
75%	54.000000
max	121.000000

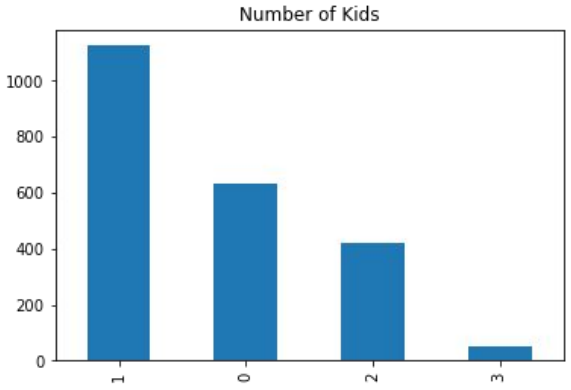


mean	51640.236126
std	20601.760369
min	1730.000000
25%	35434.750000
50%	51566.000000
75%	68118.000000
max	113734.000000

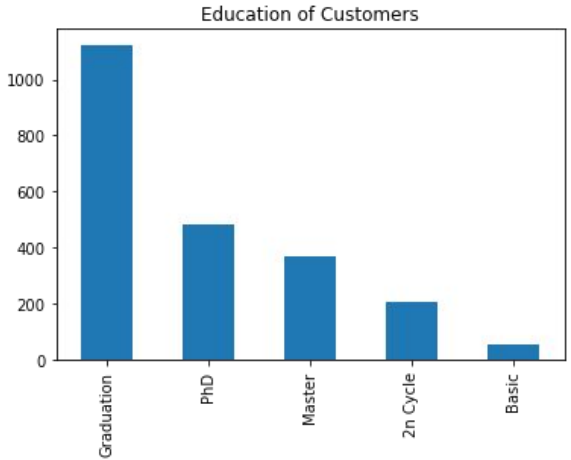
Our business currently appears to appeal more to the age group 35-54 years old, and the income group of 35,000-68,000.

Exploratory Data Analysis

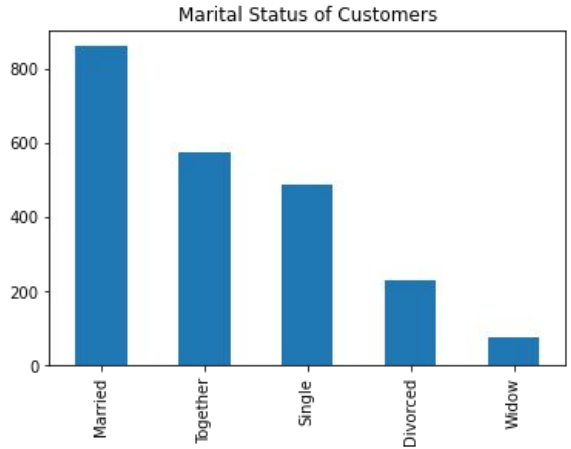
Of demographic variables



Our typical customer has 0 or 1 kid/teenager at their homes.

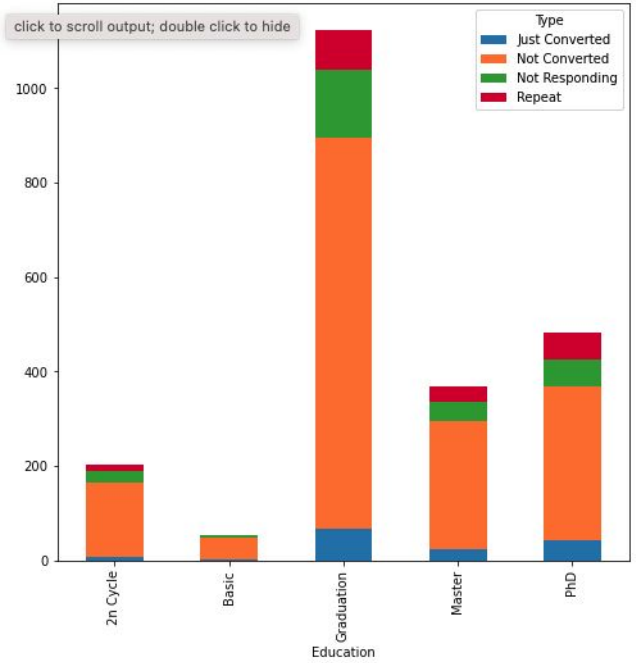


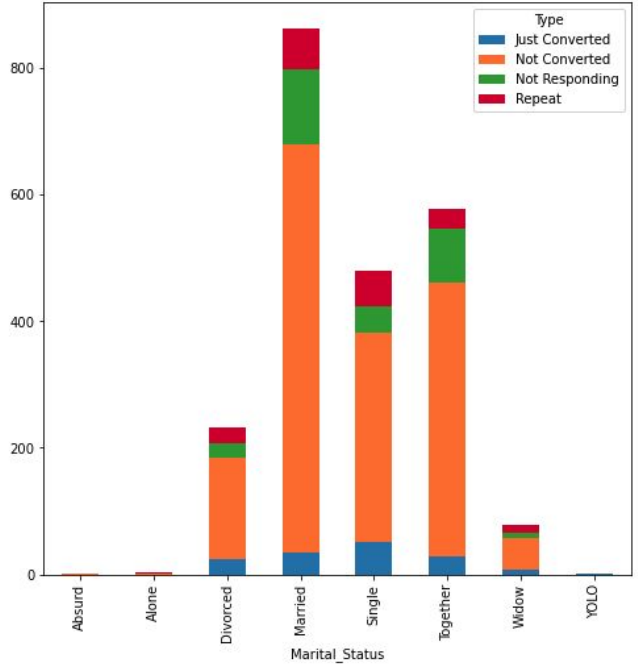
Our customers typically are college educated, a significant proportion continue higher education.



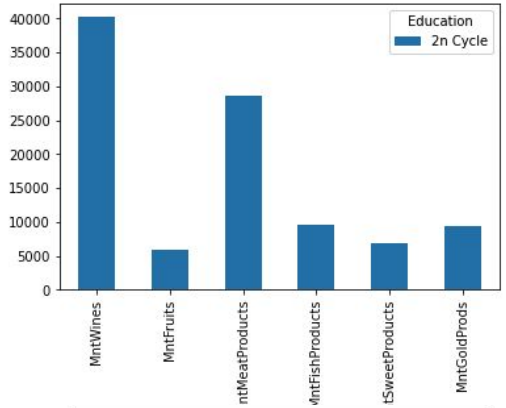
Most of our customers are married or living together with a partner.

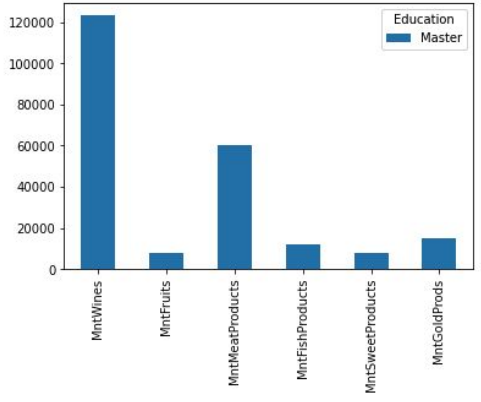
Education and Marital Status Distribution

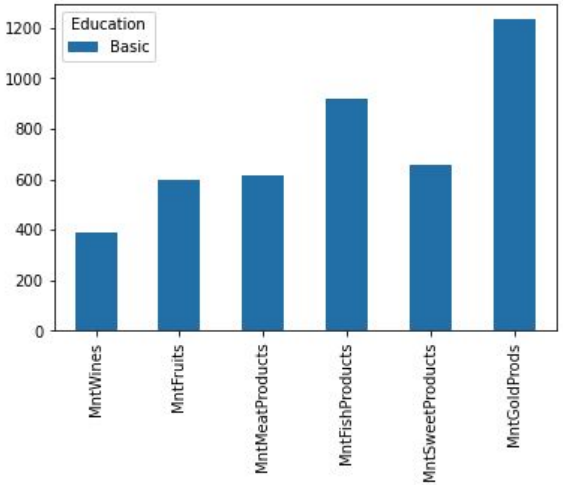


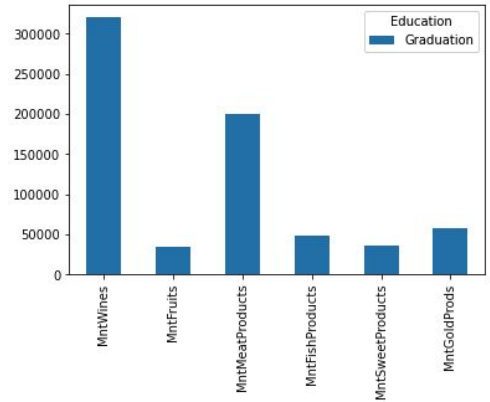


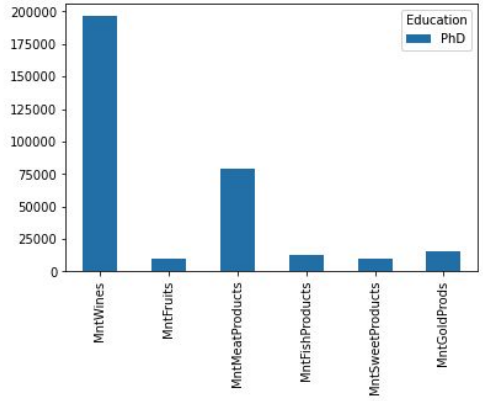
Education Distribution





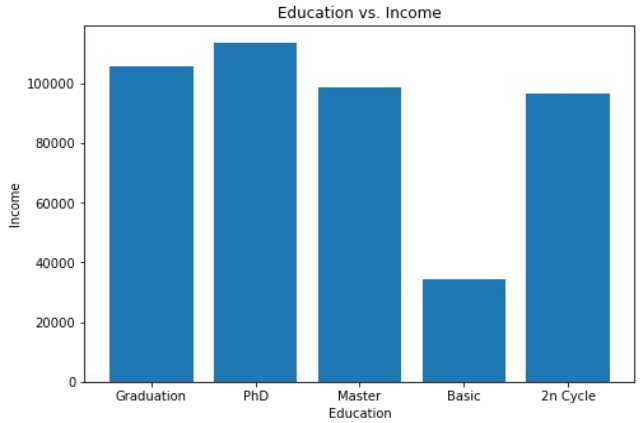


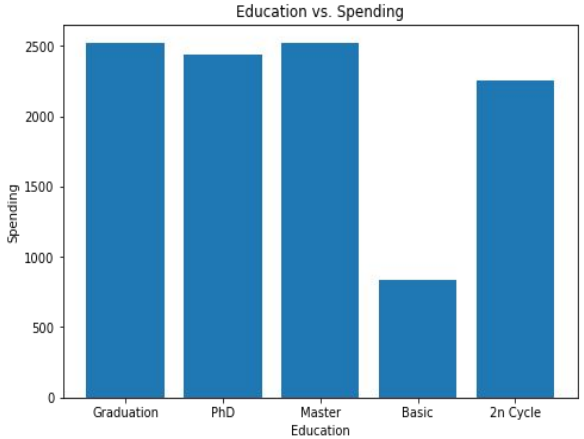


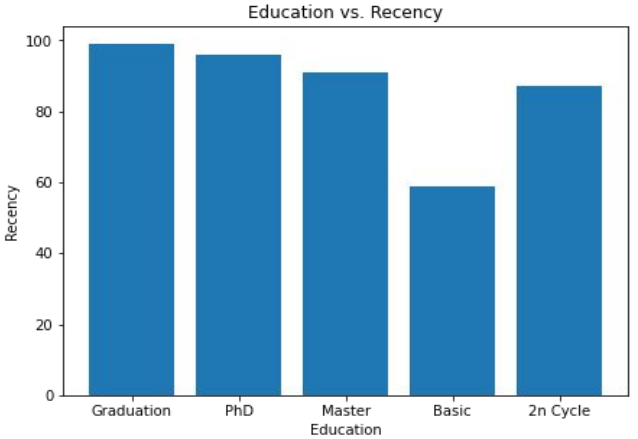


Exploratory Data Analysis

Of demographic variables







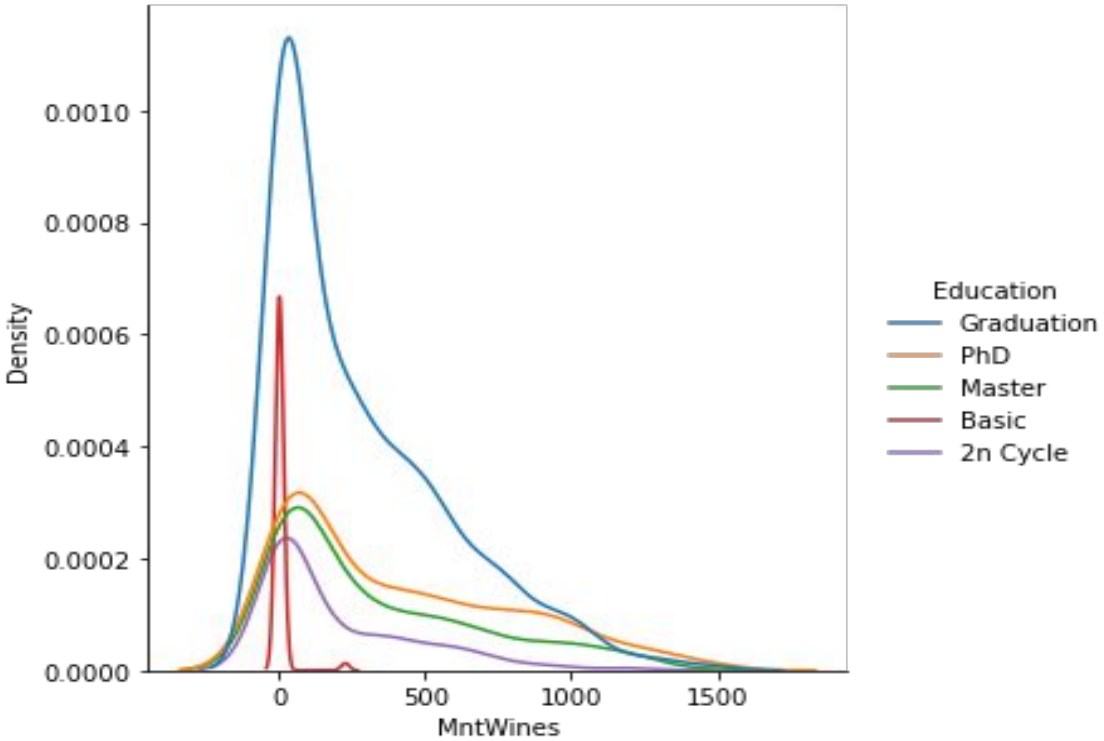
Customers with a PhD have a higher income

Customers who have a degree spend more than people without a degree

Customers in the Graduation category have a higher recency rate; buying items more recently

Exploratory Data Analysis

Of demographic variables



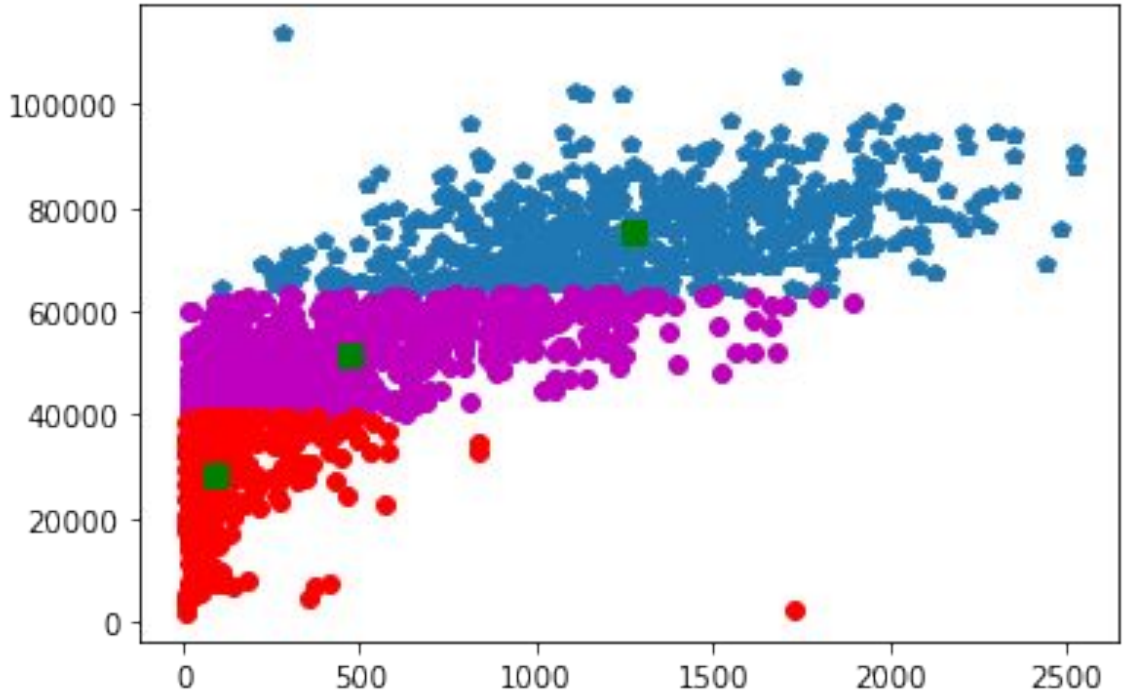
Most sold product is MntWines

Used density plot to observe the distribution of MntWines bought and education level

Graduation education level has the highest concentration of MntWines bought

K-means Clustering

Of demographic variables

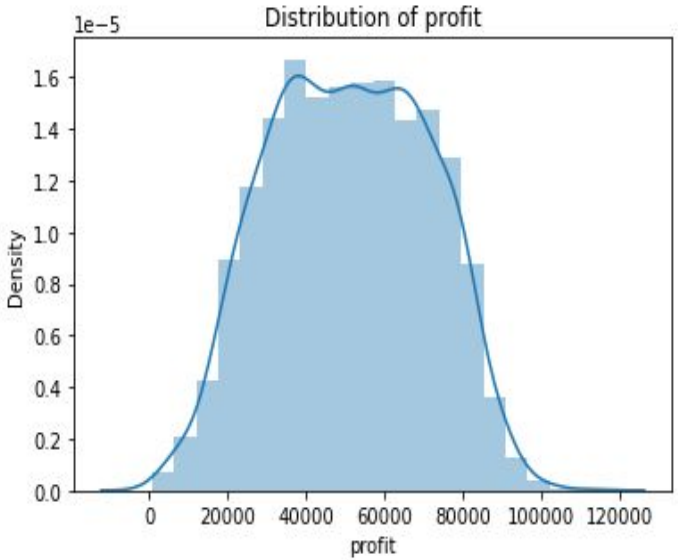


Clustering spending, and, income, education

Customers who have a higher education earn a higher income and spend more while customers who earn a lower income and do not spend as much

Profit Distribution

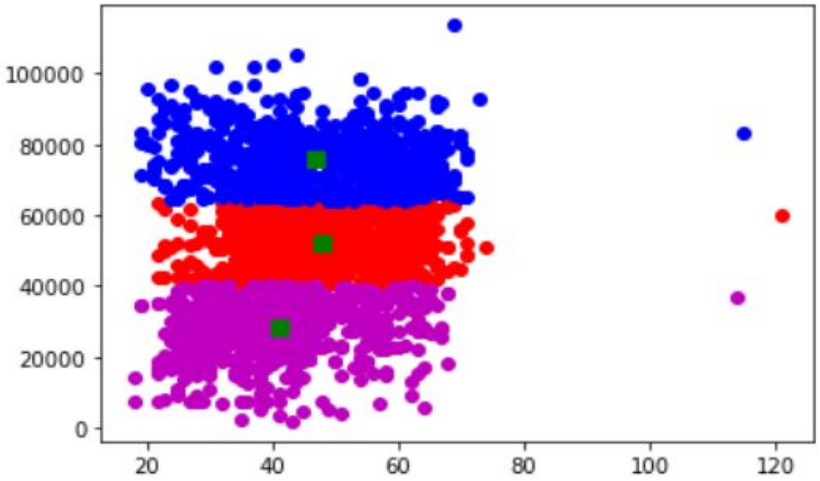
Out[56]: 'Based on the histogram of the profit destribution, we learned that the profit has a normal destribution. Also, between 20,000 and 80,000 it maintains steady'



	Adulthome	Household_Sz	profit
count	2232.000000	2232.000000	2232.000000
mean	1.643817	2.595430	51034.639800
std	0.478977	0.907085	20111.522782
min	1.000000	1.000000	717.000000
25%	1.000000	2.000000	35390.500000
50%	2.000000	3.000000	51080.500000
75%	2.000000	3.000000	66946.500000
max	2.000000	5.000000	113457.000000

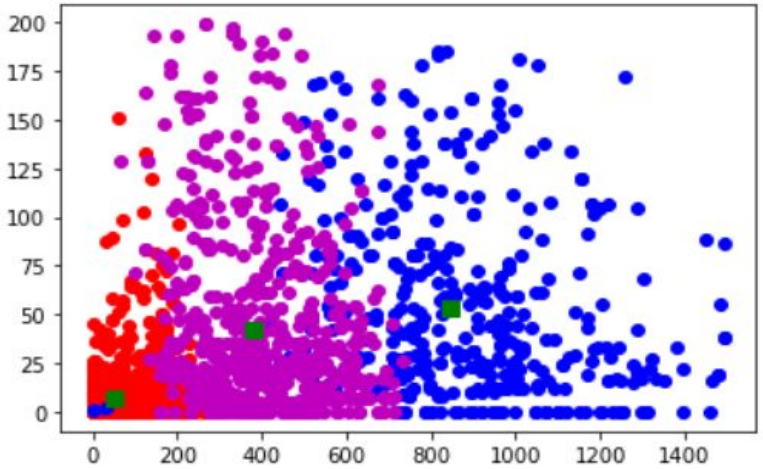
K-means Clustering

Based on demographics:



'Age', 'Income', 'Spending', 'Kidhome', 'Teenhome'

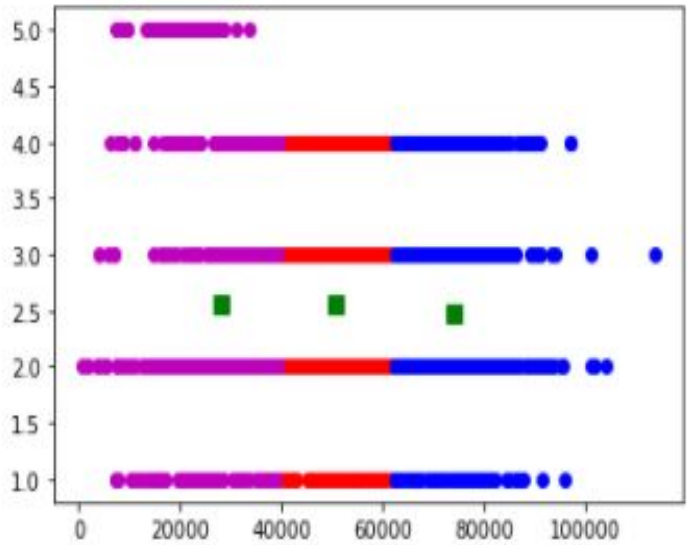
Based on purchase history:



'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth'

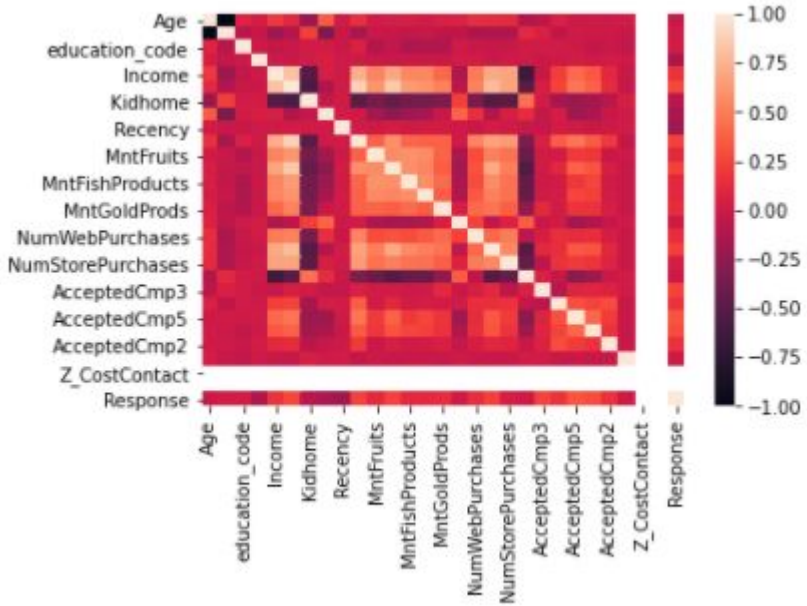
Segregation based on demographic is clearer. There is significant overlapping of clustering based on purchase history.

K-Means Clustering of Profit, Education, & Marital Status



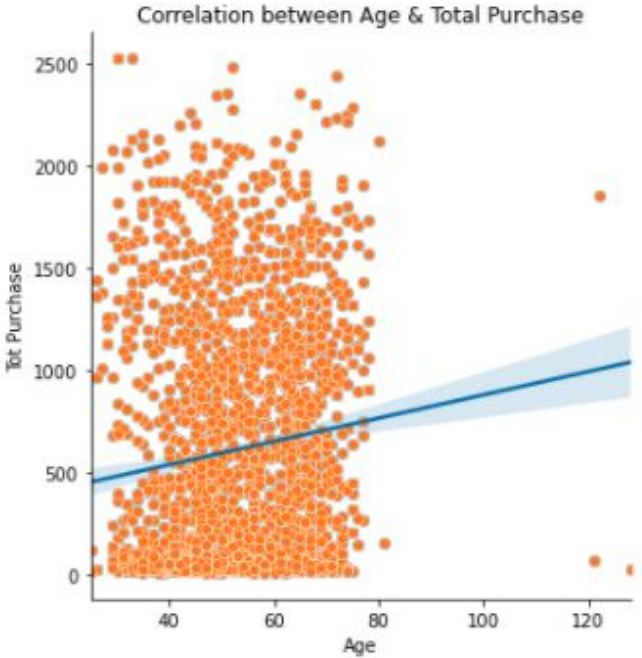
'The plotter graph indicates that the profit, education, and marital variables are not overlaping meaning that they are signifi
cant apart'

Correlation Between Variables



- Heatmap for all variables
- But does not explain further in details

Correlation Between Variables



Chi-Squared

- To determine if there is a relationship between Education and Conversion Rate (Type)

Type	Just Converted	Not Converted	Not Responding	Repeat
2n Cycle	8	157	24	14
Basic	1	47	5	1
Graduation	68	826	146	884
Master	25	269	43	32
PhD	44	324	57	57

Chi-Squared

	Value
Degree of Freedom	1
Chi Square	13.508644118
P Value	0.0002374670
Alpha	0.05
Critical Value	3.8414588206

Statistical Hypothesis Test:

H0: There is no relationship between Education and Conversion Rate

HA: There is a relationship between Education and Conversion Rate

Due to the p value is less than alpha value:

Reject H0; There is a relationship between Education and Conversion Rate

Predictive Models

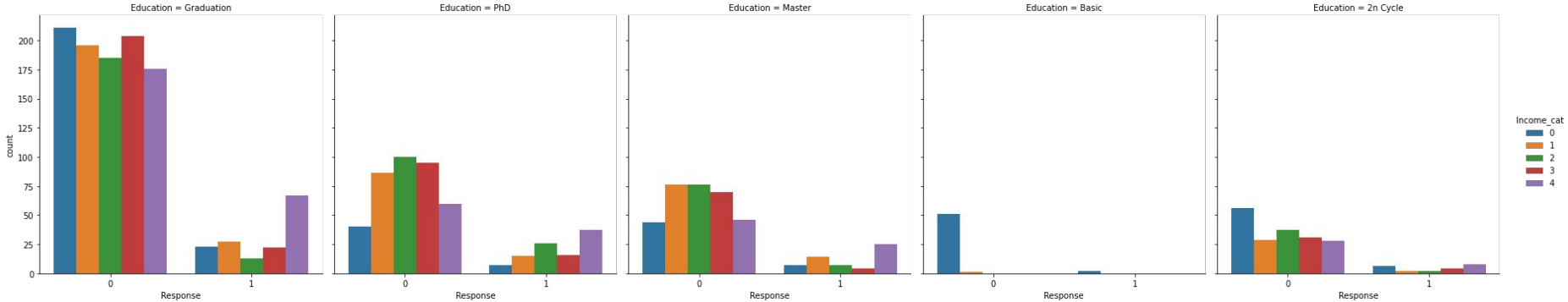
To maximize campaign profit

Responses to Marketing Campaigns

- There are 5 preceding marketing campaigns, to which a customer may/may not respond (by making a purchase).
- The customers' responses are stored in binary variables: 1 for yes an 0 for no.
- Our task is to predict whether or not a customer will respond to the next (6th) campaign.
- If we manage to build an effective model to predict responses, we can target people who are more likely to respond, hence saving marketing cost and maximizing revenue.

Exploratory Data Analysis

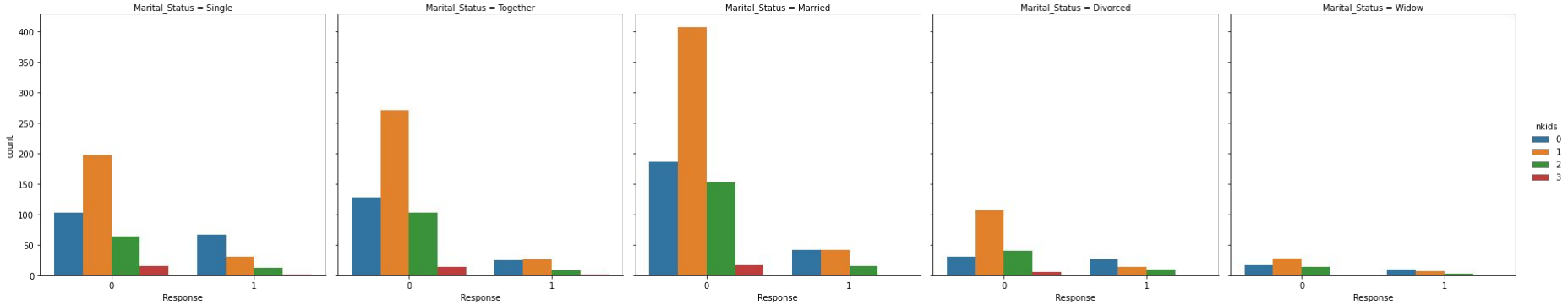
In relation to response to 6th campaign



the highest rejection percentage is apparent on the education level = "Graduation". However for this education level, income category 4 is more likely to response to the campaign. Income **category 4 (high income)** with education level of **Graduation, PhD and Master** are more likely to respond to the campaign.

Exploratory Data Analysis

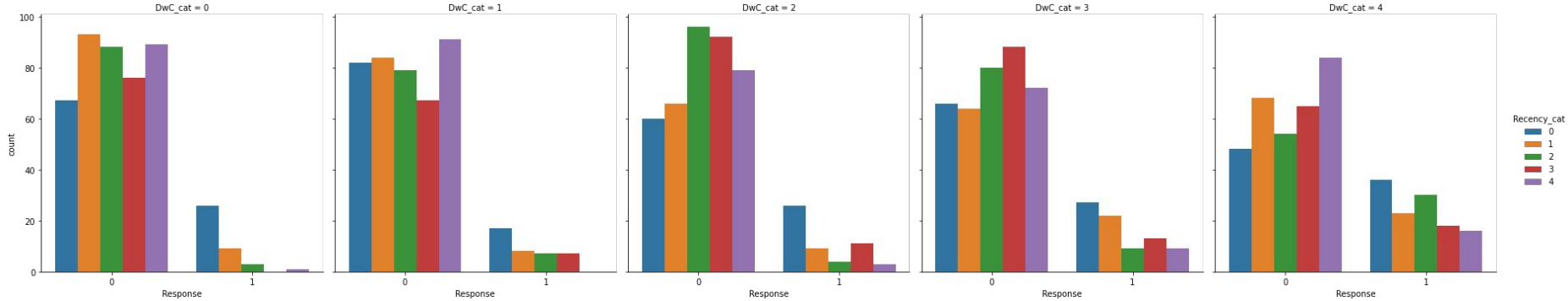
In relation to response to 6th campaign



Across all Marital_Status, people with 0 kids are more likely to respond.
Highest rejection rate is apparent from people with 1 kids, across all Marital_Status.

Exploratory Data Analysis

In relation to response to 6th campaign



The **longer** a customer has been a customer for the company, the more likely they are to respond to the 6th campaign
Across all stages of customer account's life within the company, customers that **purchased recently** are more likely to response to the 6th campaign.

Preparation for Machine Learning

To predict response to 6th campaign

Independent variables to include (avoid leakage variables):
'Age', 'education_code', ‘Marital_Status', 'Income', 'Spending', 'Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Days_with_company'.

Dependent variable:
‘Response’ to 6th campaign, 1 is yes 0 is no response.

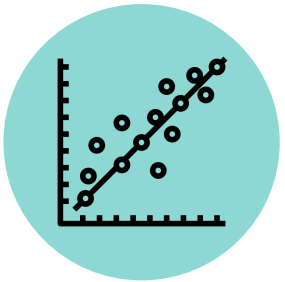
Further steps:

1. Make sure categorical variables are properly coded
2. Create dummy variables where necessary (where categorical variables are not ordinal).
3. Separate to training and testing sets
4. Create k-folds validation
5. Build machine learning algorithms

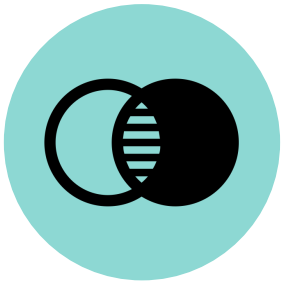
Machine Learning Algorithms

To predict customer's response to 6th campaign

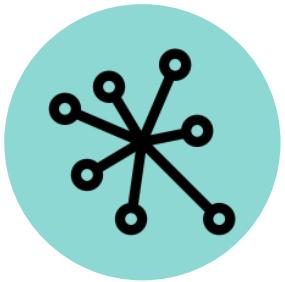
Logistic Regression



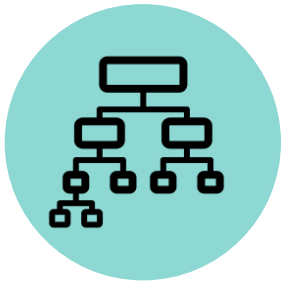
Linear Discriminant



Naive Bayes



K-means



Decision Tree

SVM

Machine Learning Algorithms

To predict customer's response to 6th campaign

Logistic Regression

Confusion Matrix

		Pred		
		0	1	Tot
Actual	0	369	17	386
	1	50	11	61
	Tot	419	28	447

Accuracy = 85% (correctly classified)
Precision = 39% (correct positives)

No Model: Cost = 447 * 3 = 1341 Rev = 61 * 11 = 671 Profit/Loss = ($670)	Send campaign to everyone, see who responses.

With Model: Cost = 28 * 3 = 84 Rev = 11 * 11 = 121 Profit/Loss = 37	Send campaign to customers who are predicted to response by model.

Machine Learning Algorithms

To predict customer's response to 6th campaign

			No Model			Model		
Model	**Accuracy**	**Precision**	**Cost ($)**	**Rev ($)**	**Profit/Loss ($)**	**Cost ($)**	**Rev ($)**	**Profit/Loss ($)**
LR	0.850112	0.392857	1341	671	-670	84	121	37
LDA	0.89038	0.65	1341	671	-670	120	286	166
KNN	0.834452	0.314286	1341	671	-670	105	121	16
DT	0.856823	0.477612	1341	671	-670	201	352	151
Naïve Bayes	0.756152	0.273585	1341	671	-670	318	319	1
SVM	0.863535	N/A	1341	671	-670	0	0	0

Thank you!