Identifying the successful drivers

Mansoureh Aghabeig

Main steps

- Understanding data
- Preparing data for applying machine learning model
- Analysis data
- Applying machine learning model
- Steps for implementing the machine learning model
- Choosing machine learning model
- Suggestion for future work

Understanding data

Understanding data

Checking data from different aspect:

- Determining different attributes (columns) in data
- Type of data to be sure it is consistent with model we are going to apply
- Missing values in data

Statistical information of data

```
[ ]: import pandas as pd
     import matplotlib.pyplot as plt


     df = pd.read_csv ('TV_CASE.csv', sep=';')
     print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5347 entries, 0 to 5346
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Programme           5347 non-null   int64
 1   Station             5347 non-null   int64
 2   datetime            5347 non-null   object
 3   cost                5347 non-null   float64
 4   profit              5332 non-null   float64
 5   order_sum           5332 non-null   float64
 6   male in %           5347 non-null   float64
 7   new customer in %   5347 non-null   float64
 8   vis                 5347 non-null   int64
dtypes: float64(5), int64(3), object(1)
memory usage: 376.1+ KB
None
```

We can see :

- 9 attributes in data
- Total number of data is: 5347 rows
- Two attributes contain missing values
- We have one non-numerical type in data

Checking missing values

```
[ ]: df.isnull().any()

[ ]: Programme            False
     Station              False
     datetime             False
     cost                 False
     profit                True
     order_sum             True
     male in %            False
     new customer in %    False
     vis                  False
     dtype: bool
```

```
[ ]: df.isnull().sum()

[ ]: Programme             0
     Station               0
     datetime              0
     cost                  0
     profit               15
     order_sum            15
     male in %             0
     new customer in %     0
     vis                   0
     dtype: int64
```

- Checking better data to understand where is the missing values
- The col **profit and order_sum** contain missing values.
- The total number of missing values are **15 rows**

Preparing data for applying machine learning model

Handling missing values

It is impossible to apply a machine learning method on dataset contains missing values.

There are two main solutions for handling missing values:

- Deleting rows contain missing values
- Imputing (replacing) missing values with a value

As the number of rows contain missing values in our data is so low in compared to whole data set, exactly 0.2%, we decide to simply remove rows containing missing values

Analysis data

Success metric

We should define which attribute in data is the target one. In other word we should decide if one driver is a successful based on this attribute.

In our date, there are four attributes which may show the trend of success in one driver:

- **Profit:** Predicted profit provided by the data science team
- **Order_sum:** Predicted orders provided by the data science team
- **New customer in:** Predicted new customers in % who saw the spot provided by the data science team
- **Vis:** Predicted visits on the Mister Spex website driven by the spot (provided by the data science team)

What is the main success metric?

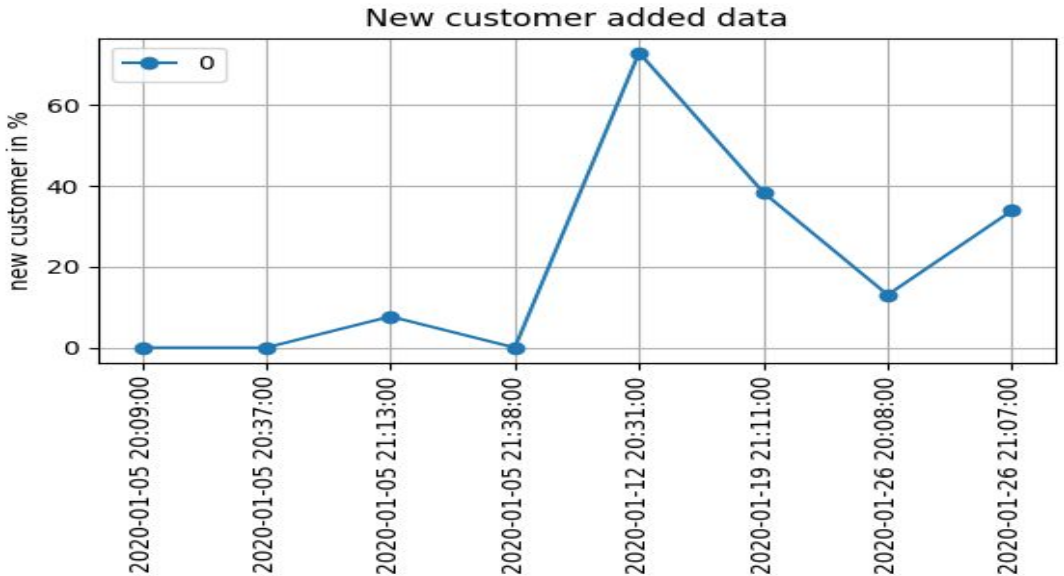
- We know that if each of these four metrics have an increasing trend during time, can expect the successfulness in a driver.
- But we need to choose one or two of these metrics for building our model.
- But which of them?

Solution for finding the best metric

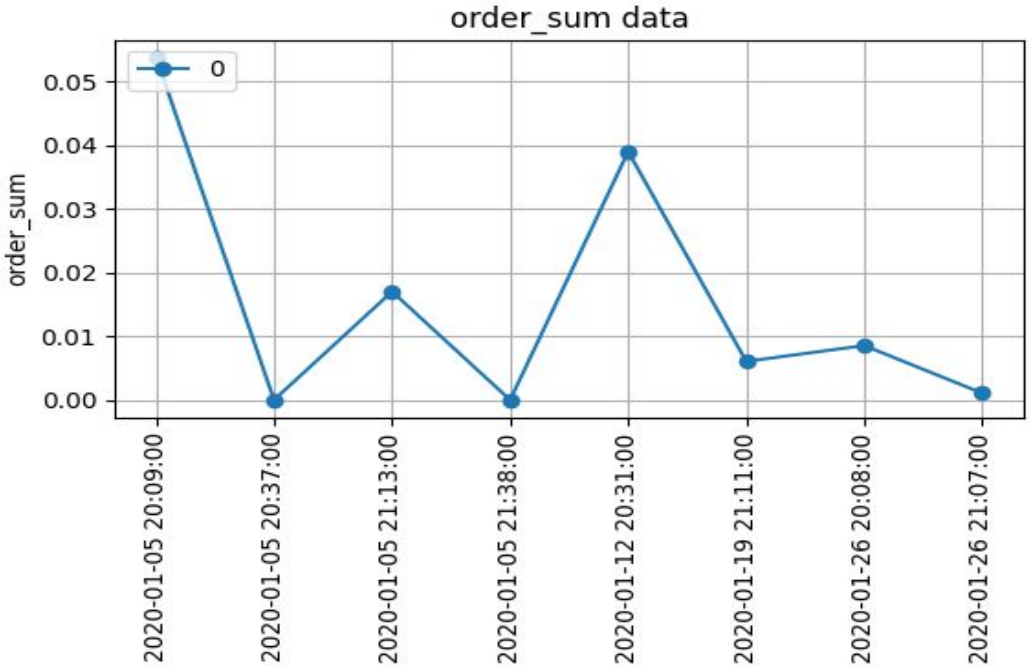
- Hypothesis 1 :
  - One solution is plotting time series data for each driver based on these four metrics.
  - We did it for example for “Programme 0”
  - We can see time series plots with different shapes for each metrics
  - Also we can see different pick during time for each metric

- Interpretation:
  - We cannot simply say that which metric is the best one by seeing their time_series data
  - Because for example, for “programme 0”, in first days of month, the order sum is quite high while the profit value is low. Or you may see in the mid of the month, the new customer added is increased significantly while the profit stay static.

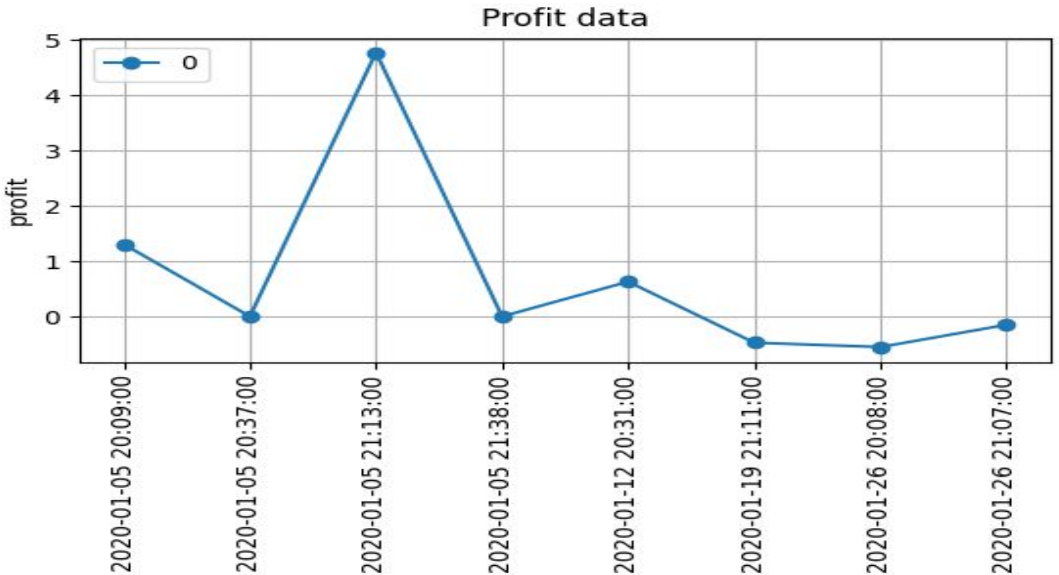
Time series plot for Program 0- New customer



Time series plot for Program 0- Order_sum



Time series plot for Program 0- Profit



Time series plots for Program 0-Visit

Solution for finding the best metric

- Hypothesis 2:
  - We should find what is the relationship between different metric
  - For this, we need to find the correlation values between metrics
  - Please note, a correlation value above zero shows a positive relation between two metrics, in other word by increasing the one, the other increase also and a correlation below zero shows an inverse relation between two metrics

Correlation values between metrics

Metric	Profit	Visit	Order_sum
Profit	1	----------	----------
Visit	0.051	1	----------
Order_sum	0.23	0.3	1
New_customer	-0.02	-0.02	-0.007

Solution for finding the best metric

Interpretation for Hypothesis 2:

- We can see there is a **positive relation among all metrics except new_customer**.
    - This means as profit, order_sum and number of visit are increased we have decrease in new customer
    - But comparing the correlation value of new customer with other metrics shows very minor relationship

Solution for finding the best metric

Final decision about successful metric:

- Among three metrics: “**Profit, Sum_order & Visit**” as they have all the positive correlation with each other, we can consider each of them as success metric
- In this study, we consider **profit** as it has a more meaning for customer.
- About metric new customer as both it has negative and also very low relation with other metrics we can ignore this metric as a successful one

Applying machine learning model

Choosing a data science model

Goal:

- **In this project, we want to identify each driver (program) is successful or not.**
- **There are several approaches for reaching this goal.**

Our approach:

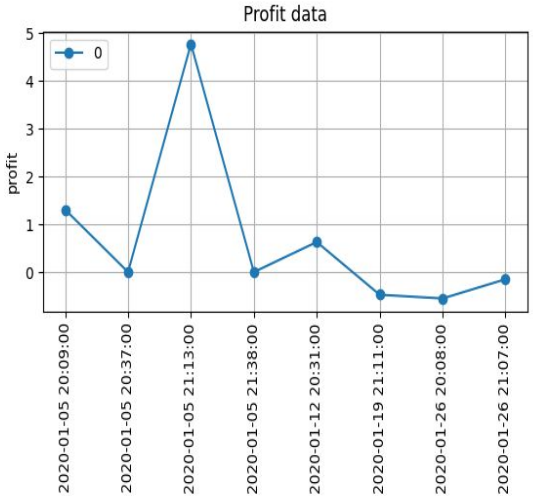
- **Predicting profit values for the other month for each driver**

Our approach

If we can find the pattern of time series of profit for this month, maybe we can predict the next time series of profit for the other month too.

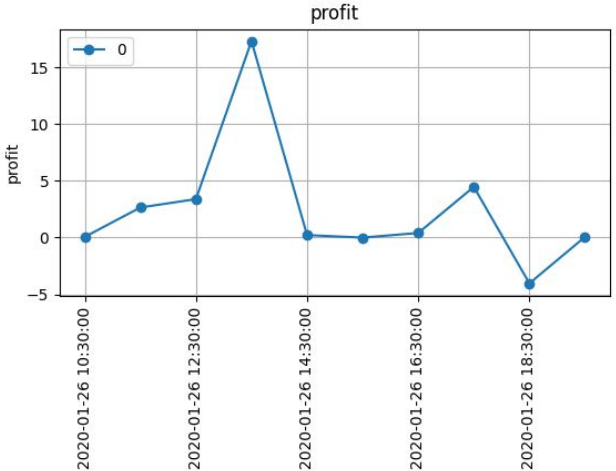
In this way, we can compare the total profit of each driver from one month to other month, and if we see an increase in total profit for the other month, we may conclude that this driver has a successful trend and we consider it as a good one.

Our approach



Machine learning Model

Predict for other month



Total Profit

N

Compute M- N

- If it is positive = Success
- If it is negative = Unsuccess

Total Profit

M

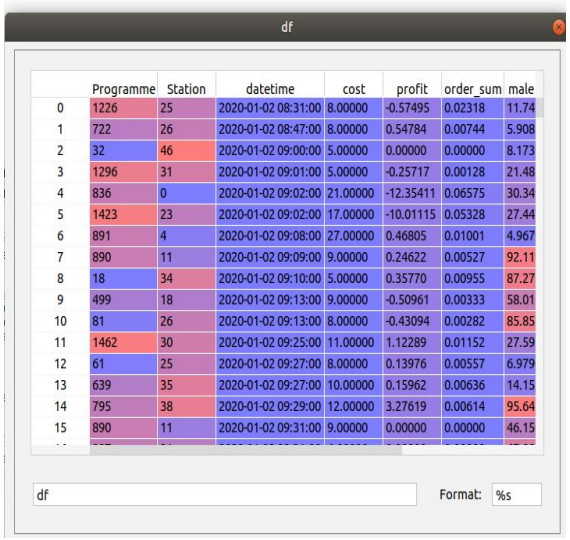
Steps for implementing the machine learning model

Generate a test data set

We need to generate a test data set which has the same statistical properties with the original dataset

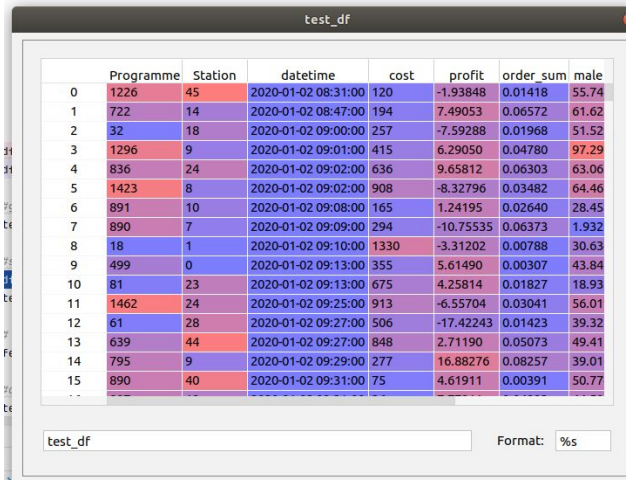
In this way, we are able to compare profit result of this month and the other month

Generate a test data set



Keeping average and standard deviation for each column

Generate Test dataset



Original dataset

Test dataset

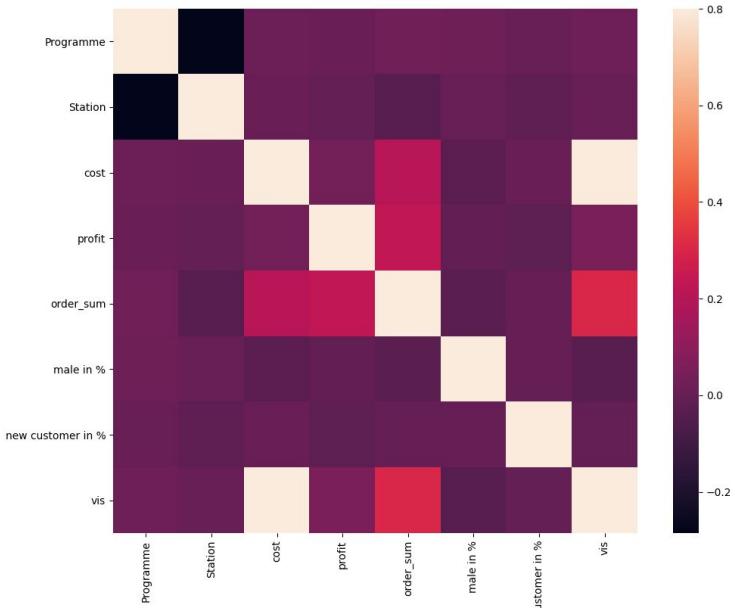
Feature selection

We need to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

We use two methods for feature (attribute) selection:

- Correlation matrix generation for Profit attribute
- Determining feature importance based on the used machine learning model

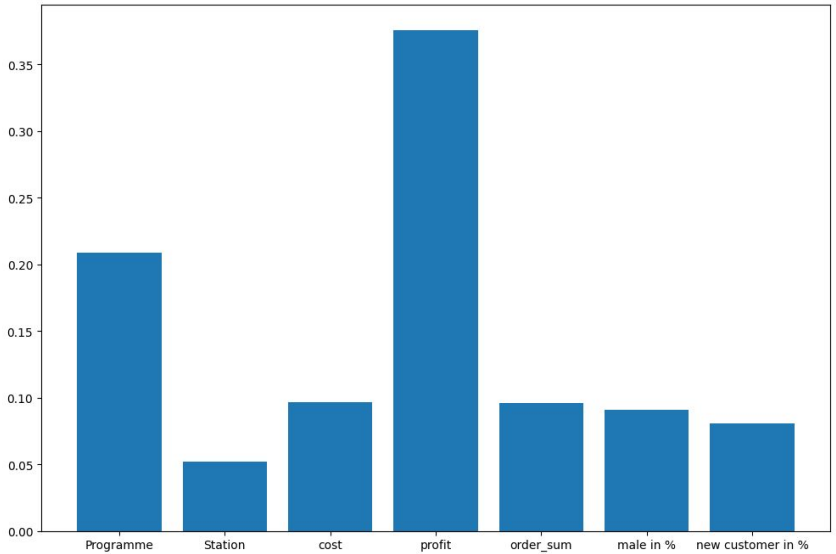
Confusion matrix for Profit



We can observe that:

- There is a high relation between profit and order_sum, cost & vis
- In other hand, there is a low relation between profit and station & new customer

Feature importance based on model



We determine the importance of each attribute based on the **machine learning model** we are using:

We can observe that for our model, **program, profit, cost, order_sum and male** have important influence

While, **station and new customer** do not have a significant influence on our model

Feature selection

Based on confusion matrix for Profile and also feature importance for our model, we decide to not consider these two attribute in our model

- Station
- New customer

Choosing machine learning model

Random forest regressor model

- It is not easy to choose an exact predictive model for the project

- We try to implement the “Random forest regressor” which applied successfully in the project “ Sales prediction over time”. As there are some common points between two project

Steps for implementing model

- Removing the Station and New customer attribute from our data set
- Dividing our whole dataset to two train and test part
- Training model based on train part of data
- We may use test part for determining the accuracy of our model (*in this version, as we just applied one model, we did not evaluate the accuracy of it, but in next version, we need to check accuracy for different models, to find the best one*)
- Applying the trained model to test data set.
- Generate the result data set

The result file

df_result

	Programme	Profit_first_month	Profit_second_month	successful
300	0	0.53294	7.08388	0.00000
845	1	-12.45779	5.23311	0.00000
1058	2	0.00000	0.85595	0.00000
449	3	2.94444	4.39006	0.00000
293	4	0.00621	5.61554	0.00000
960	5	0.20064	0.26175	0.00000
665	6	26.21667	5.18383	1.00000
555	7	65.43882	72.02362	0.00000
1281	8	-0.10729	0.30714	0.00000
1244	9	0.00000	3.31906	0.00000
1422	10	1.93976	0.42812	1.00000
180	11	0.00000	3.42333	0.00000
501	12	1.86760	2.88102	0.00000
1290	13	3.91910	6.05723	0.00000
866	14	-16.92593	-1.78538	0.00000
89	15	-6.36162	0.25532	0.00000
995	16	0.30880	0.11412	1.00000

df_result Format: %s

Suggestion for future work

Suggestion

In this project, we just apply “Random forest regressor” for train our model, but in future we may apply more models used for forecasting time series data such as

- Linear regression
- K nearest neighbor
- Stochastic gradient descent
- Decision tree
- Neural network
- Autoregressive–moving-average model

It is important to know, we have a time series forecasting model here.