

DATE

31/10/2021

TEAM MEMBERS:

1. Yashwant Tahiliani

2. Sachin D Shelke

3. Pooja Potdar

4. Pranit Pawar

5. Purva Gaikwad

6. Mohit Jain

Mentor: Divyanshu

GROUP NO.5

Contents

[**1.** **PROBLEM STATEMENT:** 2](#_Toc86584994)

[**OBJECTIVE 1:** 2](#_Toc86584995)

[**OBJECTIVE 2:** 2](#_Toc86584996)

[**2.** **INTRODUCTION** 2](#_Toc86584997)

[**3.** **EDA & CONCERNED OUTPUT:** 3](#_Toc86584998)

[**4.** **DATA CLEANING AND DATA PREPARATION STEPS** 8](#_Toc86584999)

[5. **MODEL BUILDING STEPS AND APPROACH** 10](#_Toc86585000)

[**6.** **MODEL TESTING/ EVALUATION** 10](#_Toc86585001)

[**TOP FIVE VP+MANAGER COMBINATIONS RECOMMENDATIONS:** 11](#_Toc86585002)

[**FOR EVERY FALSE PREDICTION COST:** 12](#_Toc86585003)

## **PROBLEM STATEMENT:**

In this project, you will analyze and predict the win possibilities of deals/projects for an IT consulting company and see how the possibility of winning a deal is impacted by other variables. This will enable the IT consulting company to manage the effort

required to win a deal to meet the growth targets.

**OBJECTIVE 1:** Predictive Analytics - Build a ML model to predict the probability

of win/loss for bidding activities for a potential client.

**OBJECTIVE 2:** Prescriptive Analytics – Identify variable/s that are most likely to

help in converting an opportunity into a win.

## **INTRODUCTION**

Win prediction is a data set that contains the historical data of past deals / bids placed by various VP and managers. It also contains the cost of deal, deal date, status of deal (win/ loss), location, sector, client category & solution type.

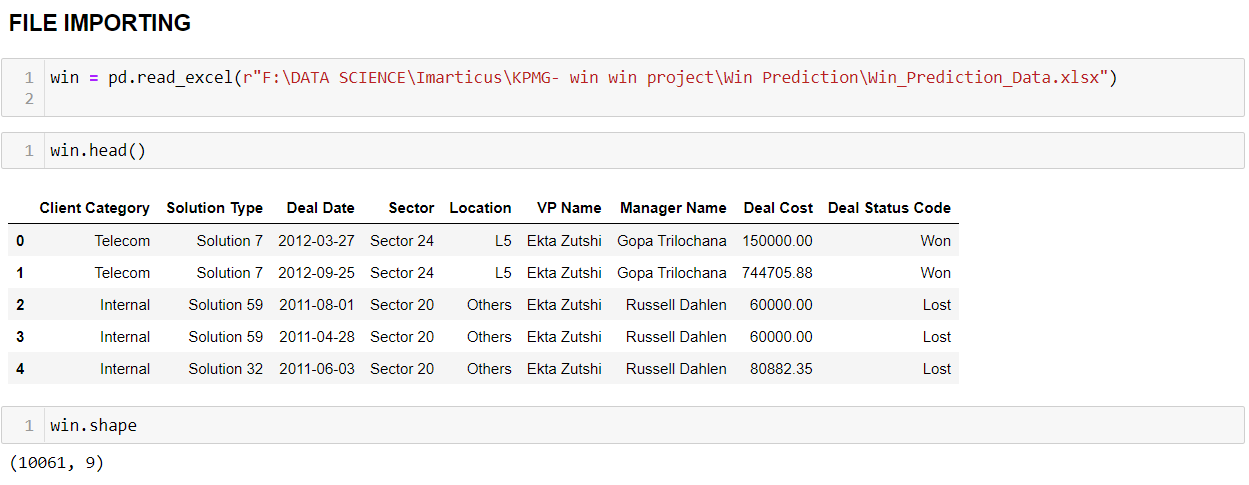
IT company needs us to make a predictive model of probable WIN deal/bid and factors affecting that win.

## **EDA & CONCERNED OUTPUT:**

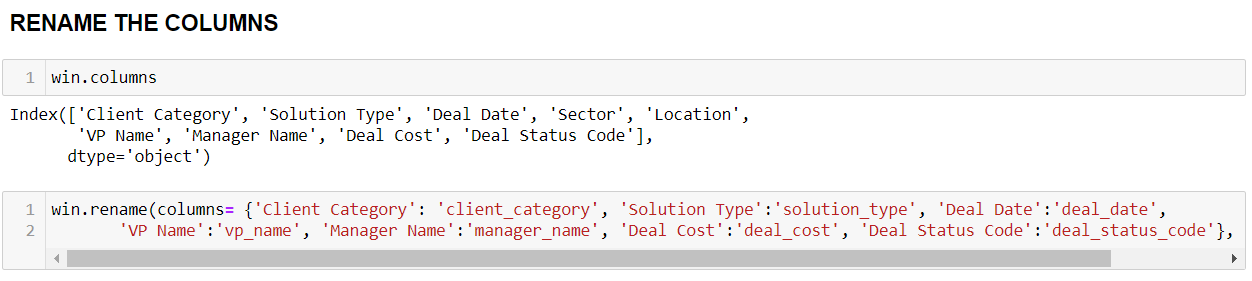
|  |  |
| --- | --- |
|  | **Bar chart:** Client category Vs deal Status.  **Insights:**  Highest count of deals: Others. |
|  | **Bar chart:** Client category Vs deal Status.  **Insights:**  Nulls: 79 entries  Replacing with “Others” |
|  | **Histogram:** Deal\_cost  **Insights:**  Heavily Right skewed.  We cannot use this data as it is heavily right skewed, hence we will try to make it in normal distribution. |
|  | **Histogram:** Natural Log of Deal cost  **Working:**  Adding a calculated field Log(deal cost)  **Insights:**  Now data is almost in normal distribution. We decided to convert the deal cost column into NATURAL LOG to try to make the data normally distributed and it worked. |
|  | **Bar chart & Area chart:**  Deal\_date VS count of deals.  **Insights:**  No co-relation found within deal\_date and deal\_status.  Using tableau, we tried to find the co-relation between winning a deal and date, analysis done on year basis, quarter basis, month basis & day basis but no pattern found. *Hence dropping deal date column.* |
| why date droped |
|  | **Bar chart:** Client category VS sum of deal cost  **Insights:**  Top 5 client category that Won and data arranged according to descending order of sum of deal cost involved. |
|  | **Bar chart:** Client\_category VS count of deal\_cost  (Filter: deal\_cost & Color: deal\_status)  **Insights:**  246 rows having 0.00 as deal\_cost.  Only 4 entries of WIN.  Initially the decision was made to drop 246 rows but to avoid losing data we manipulated data in following way.  While log transformation, Zeros were big problem (as log(0)= -inf), hence we added 1 to all deal cost column and then took log (Log(deal\_cost+1)).  This helped us in two things:   * No rows were lost * Zeros were handled. |
|  | **Bar Chart:** Location Vs Count of Deal Cost  **Insights:**  Locations L10, L5, L1, Others & L3 are top 5 locations in terms of count of deals win. |
|  | **Bar Chart:** Sector Vs Count of Deal Cost  **Insights:**  Sector number 23, 2, 20, 24 and 17 are top 5 sectors in terms of count of deals win. |

## **DATA CLEANING AND DATA PREPARATION STEPS**

1. Reading the file in python under name “win” (file was in excel format)



1. Renaming column names as it contained spaces. (easy to call in python)

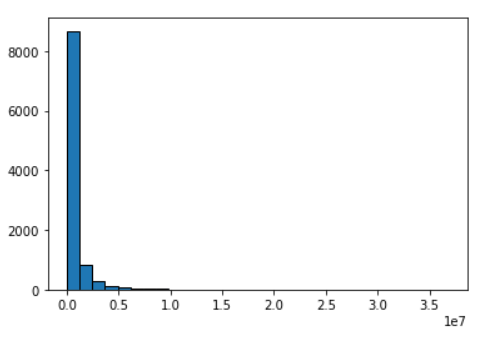
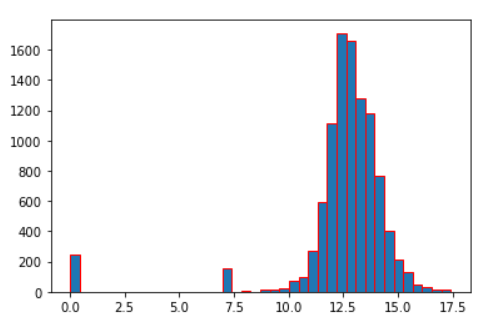


1. Replacing NULLS in Client category column by “Others” as this category has largest count.(Mode) and dropped date column.

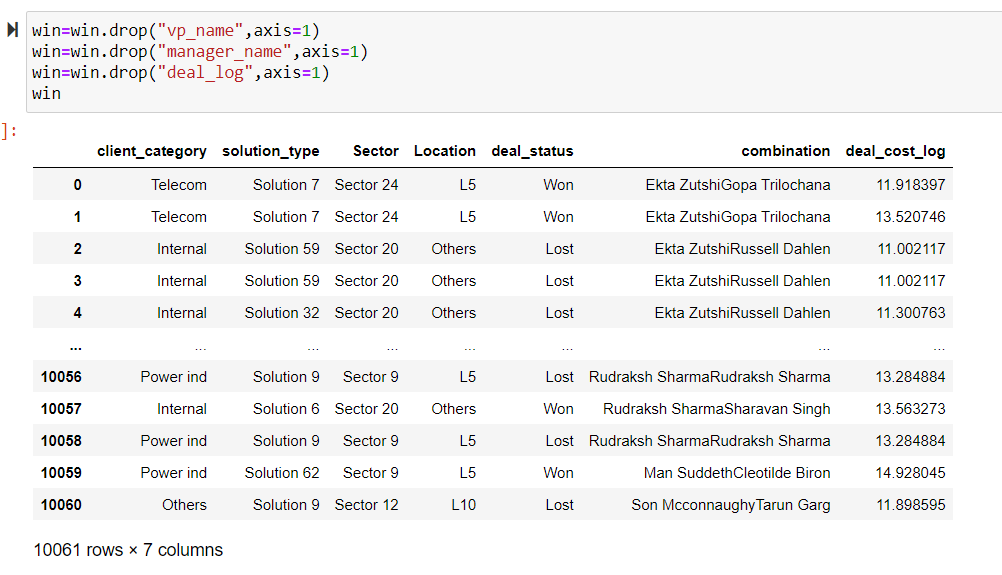
As per decided in EDA

1. Creating a new column which has combined data of VP name and Manager name. ( which will help to find out the best performing 5 combinations of vp and managers in further data analysis)

1. Natural log transformation (as mentioned in EDA) of the deal cost column and concatenating this new column in current working data frame (win).

1. Creating a data frame which contains vp’s name and manager’s name so that after transforming all the columns into numeric, we need to co relate the numeric to actual names.
2. Dropping columns vp\_name, manager\_name and deal cost. As we have combined names of VP and manager, we don’t need separate columns for that and deal cost in converted into log hence don’t need that too.



1. Converting non-numeric to numeric using Label encoder.
2. After dropping the Columns VP Name, Manager Name and Deal Log , we have combined the dataframe which contain vp name , manager’s name and their respective numeric value allotted by label encoder. After that we eliminated the duplicates and kept only unique VP –manager combinations. We will call this data frame when we will find Top 5 combination of VP and manager after creating model.

## **MODEL BUILDING STEPS AND APPROACH**

1. We decided to go with Random sampling (70%-30%). In python, while using the random sampling we used random state-555 so that the content of test and train data would remain same even if we make different models.
2. As the data is labeled, we will be performing 4 approach or 4 different models of supervised machine learning algorithms such as ,

* Logistic,
* Decision tree,
* Random Forest,
* K nearest neighbor.

## **MODEL TESTING/ EVALUATION**

Model evaluation was done by following credentials or outputs:

1. Confusion matrix
2. Accuracy score
3. Precision score
4. F1 score
5. Au-roc Score

Each model was tested by all above mentioned values so that best performing model could be selected.

As mentioned below, after comparison of all evaluation criteria for all 4 models best was selected.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Presicion | F1 Score | AuROC | Confusion  matrix |
| Logistic Regression | 0.62 | 0.005 | 0.01 | 0.52 | [[1876 1129]  [ 8 6 ]] |
| Decision Tree | 0.78 | 0.61 | 0.67 | 0.77 | [[1686 456]  [ 198 679]] |
| KNN | 0.68 | 0.49 | 0.54 | 0.66 | [[1507 575]  [ 377 560]] |
| Random Forest | 0.8 | 0.59 | 0.69 | 0.8 | [[1748 461]  [ 136 674]] |

## **TOP FIVE VP+MANAGER COMBINATIONS RECOMMENDATIONS:**

* We are going to have the **weighted probability** by multiplying win probability with deal cost as a factor of all the deal cost for a particular combination of VP and Manager Combination
* Per weighted probabilities, the top five combinations are recommended.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic | Decision Tree | KNN | Random forest |
| 1 | 623 Varsha Arora Rahul Singh | 0 Alam Syed Alam Syed | Varsha Arora & Rahul Singh | 623 Varsha Arora Rahul Singh |
| 2 | 606 Son Mcconnaughy Manish Saundriyal | 340 Mayank Mewar Aman Gupta | Saurabh Singh & Ayusha Gupta | 425 Rahul Bajpai Akshay Gosh |
| 3 | 581 Saurabh Singh Raj Pinani | 289 Mangesh K. Singh Earline Langton | Russell Dahlen & Martin Cheyne | 15 Alam Syed keshar ansari |
| 4 | 20 Ankita Aggarwal Akul Jindal | 297 Mangesh K. Singh Kaniksh Dhyani | Molly Eakes & Heath Mullarkey | 115 Ekta Zutshi Andrew Michalowski |
| 5 | 213 Hardeep Suksma ram dutt gupta | 298 Mangesh K. Singh Kendra Tripathi | Gayle Molter &  Raj Pinani | 3 Alam Syed Geoffrey Recker |

## **FOR EVERY FALSE PREDICTION COST:**

**Approach:** We are going to take an **average cost of all the incorrect predictions**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | F1 Score | AuROC | Confusion  matrix | Average Lost per deal |
| Logistic Regression | 0.62 | 0.005 | 0.01 | 0.52 | [[1876 1129]  [ 8 6 ]] | 686316.42 |
| Decision Tree | 0.78 | 0.61 | 0.67 | 0.77 | [[1686 456]  [ 198 679]] | 812471.67 |
| KNN | 0.68 | 0.49 | 0.54 | 0.66 | [[1507 575]  [ 377 560]] | 87225.88 |
| Random Forest | 0.8 | 0.59 | 0.69 | 0.8 | [[1748 461]  [ 136 674]] | 645536.01 |