IMARTICUS LEARNING

KPMG

WIN PREDICTION ANALYTICS

DATA SCIENCE PRODEGREE CAPSTONE PROJECT

DSP-09

ARUN KUMAR

DEEPALI SINGH

GAUTHAM R

MANOJ SAHU

NEHA PURANIK

TABLE OF CONTENTS

1. PROBLEM STATEMENT

IT firms compete for winning large deals by designing and proposing solutions to their clients. These deals often differ from each other in terms sector of the client, solution to be delivered, technology to be used and the scope of the project.

By predicting the probability of winning a deal, the engagement teams can prioritize the pipeline of opportunities to staff the most attractive options first. With the probability of winning known in advance, deal engagement manager can ensure that for the most profitable deals there are resources available.

OBJECTIVE

Your Organization puts in a lot of effort in bidding preparation with no indications whether it will be worth it. With multiple bid managers and SBU Heads willing to work on every opportunity, it becomes difficult for the management to decide which bid should be given to which bid manager and SBU Head. You are hired to help your organization identify the best bid manager-SBU Head combination who can convert an opportunity to win with the provided data points.

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.

2. DATA DEVELOPMENT

- The given dataset was in .XLSX format, which was converted to .CSV format and then imported to Jupyter Notebook.

```
import os
import pandas as pd
```

```
os.chdir("E:/DATA SCIENCE/Capstone Project/Win Prediction")
```

```
fullraw = pd.read_csv("Win_Prediction_Data.csv")
```

```
fullraw.head(15)
```

	Client Category	Solution Type	Deal Date	Sector	Location	VP Name	Manager Name	Deal Cost	Deal Status Code
0	Telecom	Solution 7	27-Mar-12	Sector 24	L5	Ekta Zutshi	Gopa Trilochana	150000.00	Won
1	Telecom	Solution 7	25-Sep-12	Sector 24	L5	Ekta Zutshi	Gopa Trilochana	744705.88	Won
2	Internal	Solution 59	1-Aug-11	Sector 20	Others	Ekta Zutshi	Russell Dahlen	60000.00	Lost
3	Internal	Solution 59	28-Apr-11	Sector 20	Others	Ekta Zutshi	Russell Dahlen	60000.00	Lost
4	Internal	Solution 32	3-Jun-11	Sector 20	Others	Ekta Zutshi	Russell Dahlen	80882.35	Lost
5	Internal	Solution 32	24-May-11	Sector 20	Others	Ekta Zutshi	Russell Dahlen	80882.35	Lost
6	Internal	Solution 59	3-Nov-11	Sector 2	L10	Mervin Harwood	rahul sharma	526176.47	Won
7	Govt	Solution 7	17-Sep-12	Sector 13	L5	Sargar Deep Rao	Vidur Hukle	409705.88	Lost
8	Consumer Good	Solution 42	11-Apr-12	Sector 12	L10	Lilli Storrs	Md. Daud	1032352.94	Won
9	Internal	Solution 59	17-Nov-11	Sector 20	Others	Sargar Deep Rao	Hardeep Suksma	558823.53	Lost
10	International Bank	Solution 6	11-Feb-12	Sector 2	L10	Long Bergstrom	Luv Malhotra	316176.47	Won

3. EXPLORATORY DATA ANALYSIS

Features of the given Dataset

Client Category	Solution Type	Deal Date	Sector	Location	VP Name	Manager Name	Deal Cost	Deal Status Code

Column Name	Description
Client Category	Industry in which the client works
Solution Type	The solution group the client requires
Deal Date	The date the opportunity was created
Sector	The sector for which the solution is to be provided
Location	Client location
VP Name	Sr. Manager or VP who is dealing with the client
Manager Name	Manager of the team working on the project
Deal Cost	The initial cost of the deal
Deal Status Code	Final status of the deal(won/lost)

Initial Lookup:

- 41 different Industries.
- 67 types of Solutions.
- 25 unique Sectors.
- 13 Locations.
- 43 Senior Managers.
- 278 Managers.

The given dataset contained 10,061 rows across 9 columns of different variables.

Missing Value Treatment

The dataset was almost clean, except for the 'Client Category' contained few missing values which was visualized and treated as below.

```
import numpy as np

fullraw.isnull().sum()
```

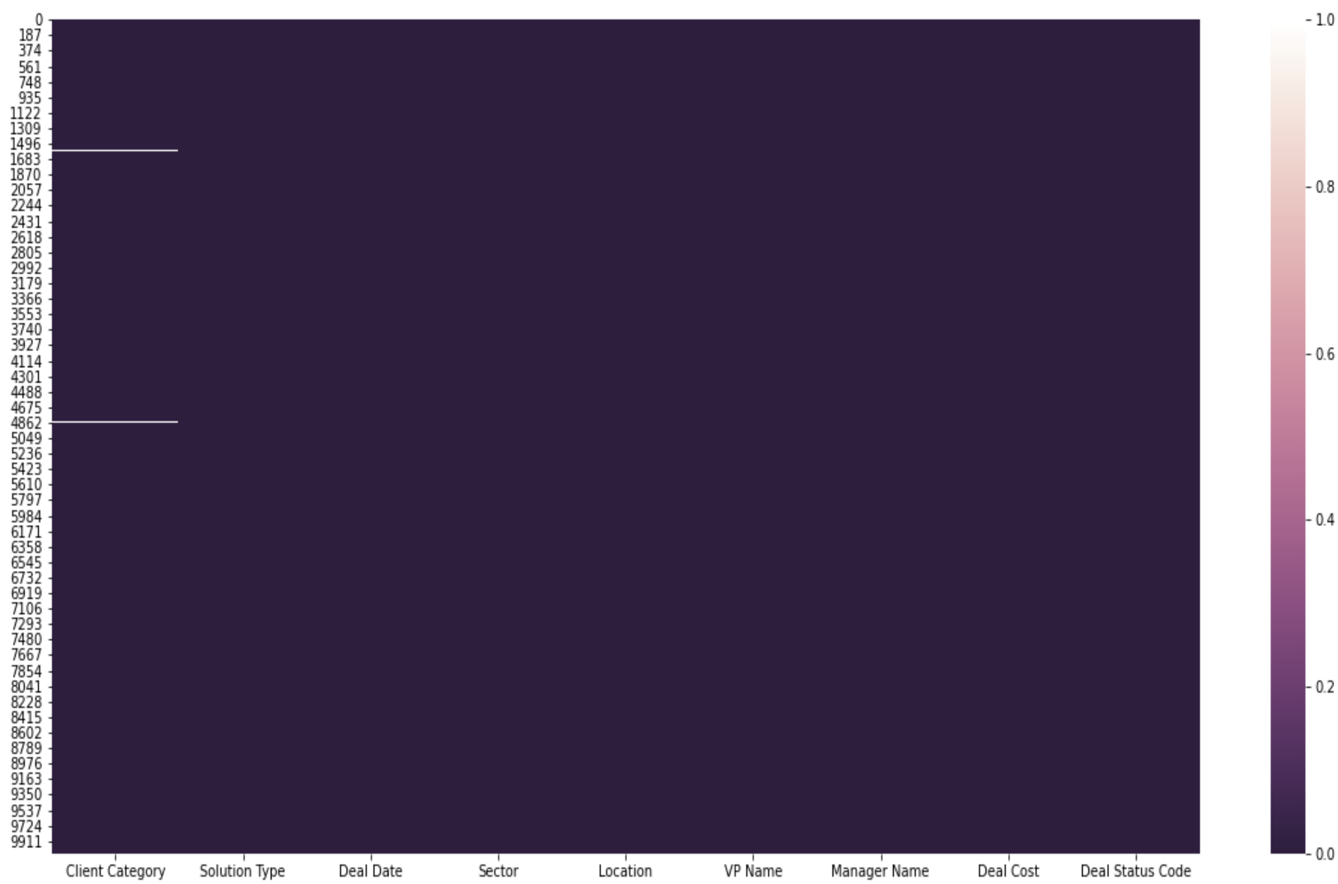
```
import seaborn as sns
import matplotlib.pyplot as plt

get_ipython().run_line_magic('matplotlib','inline')

plt.figure(figsize = (20,10))

cmap= sns.cubehelix_palette(light=1, as_cmap=True, reverse= True)

sns.heatmap(fullraw.isnull(), cmap=cmap)
```



The Client Category contained 79 missing values, for which the categories mode value was chosen for replacement. Also, a count of the distinct values of each element in the Client Category was taken into consideration.

Replacing Missing Value with Mode value.

```
fullraw['Client Category'] = fullraw['Client Category'].fillna(fullraw['Client Category'].mode()[0])
```

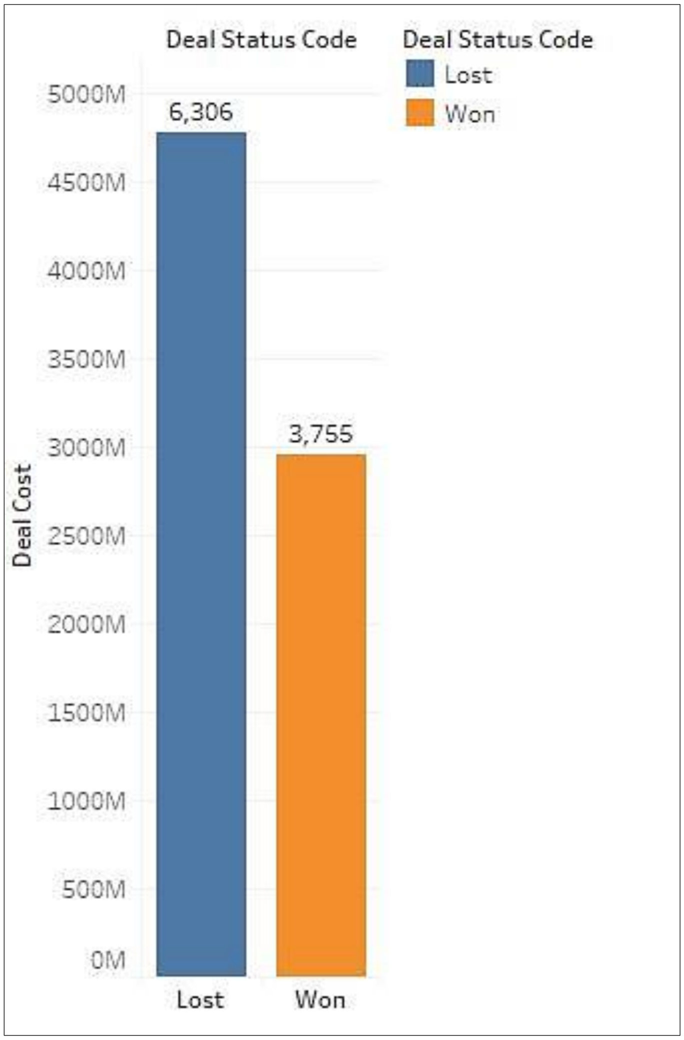
```
fullraw.isnull().sum()
```

After Treating Missing Values.

4.DATA VISUALIZATION

- **Deal Status Code**

Deal Status Code, explains whether the bid Won or Lost.



The given dataset contained

6306 Lost and 3755 Won bids.

62.8% - Lost Bids

37.2% - Win Bids.

The graph indicates the company's success rate at bidding, historically has not been a pleasant experience.

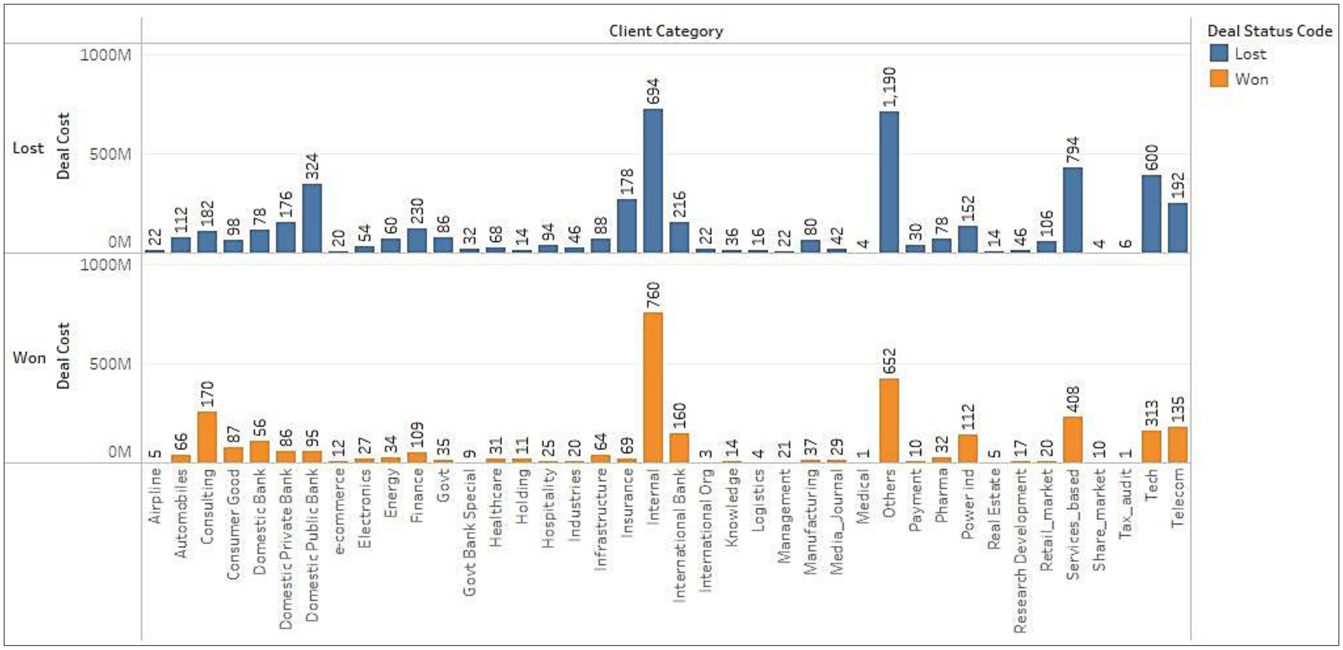
Thus, the current model becomes even more valuable at predicting the right combination of partners.

The Lost Bid forms the majority section of the given dataset, therefore the loss obtained upon every wrong prediction is about to cost much more than the rightly predicted pair.

The model's <u>False Prediction Rate</u> must be kept under check.

➢ **Client Category**

The dataset contains 41 different types of Clients.



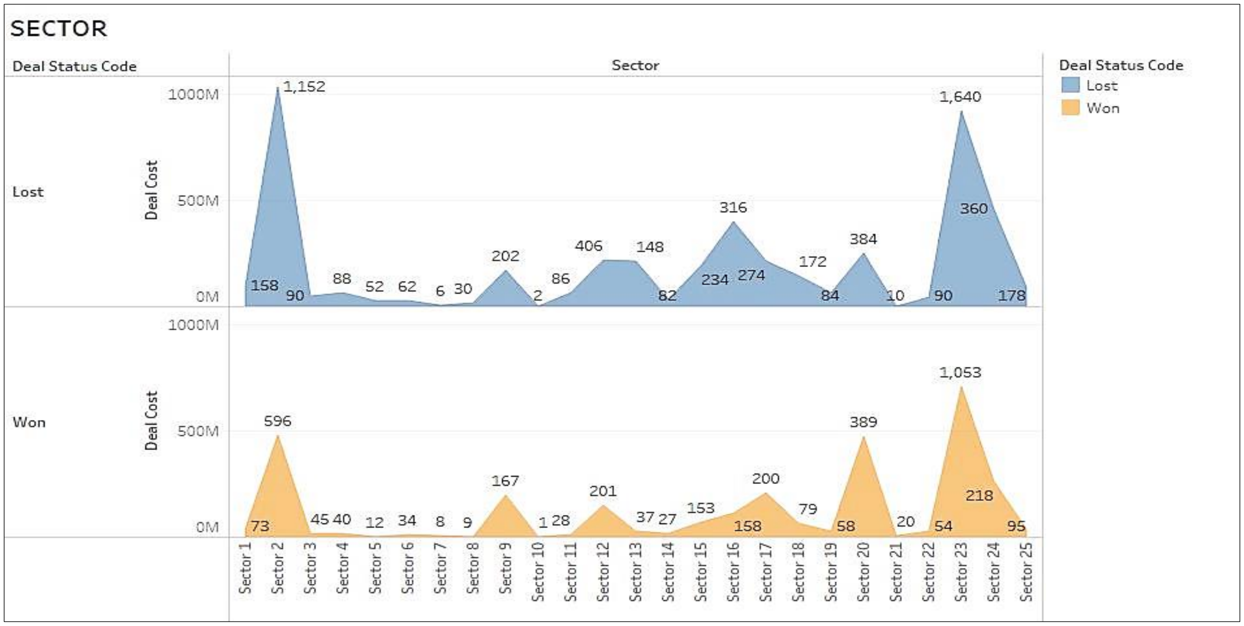
The above graph clearly shows the different categories with both Won and Lost bids in Deal Cost and the number of bids respectively.

The Top 5 Bidding Categories:

Client Category	Won	Lost	Percentage of Wins
Internal	760	694	52.27 %
Others	632	1190	35.84 %
Services_based	408	794	33.94 %
Tech	313	600	34.28 %
Domestic Public Bank	95	324	22.67 %

➢ **Sector**

The company has bid in 25 Sectors in past eight years.



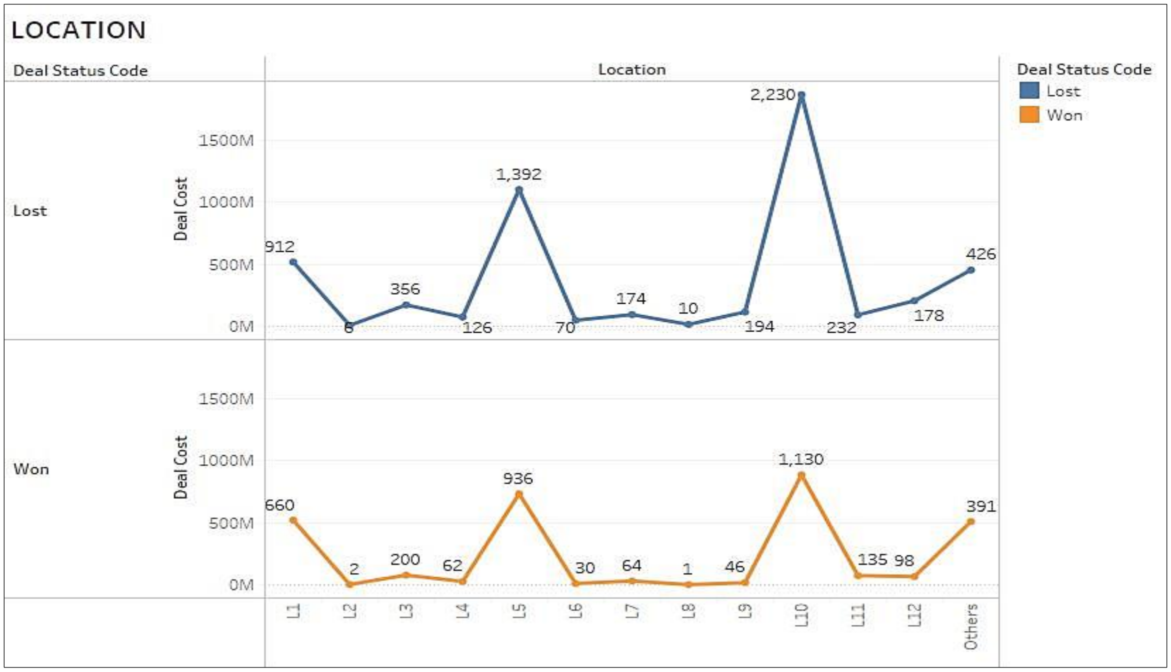
The above graph clearly shows the different sectors with both Won and Lost bids in Deal Cost and the number of bids respectively.

The Top 5 Bidding Sectors:

Sectors	Won	Lost	Percentage of Wins
Sector 23	1053	1640	39.1 %
Sector 2	596	1152	34.09 %
Sector 20	389	384	50.32 %
Sector 12	406	201	66.88 %
Sector 17	200	274	42.19 %

➢ Location

The company has bid at more than 13 Locations.



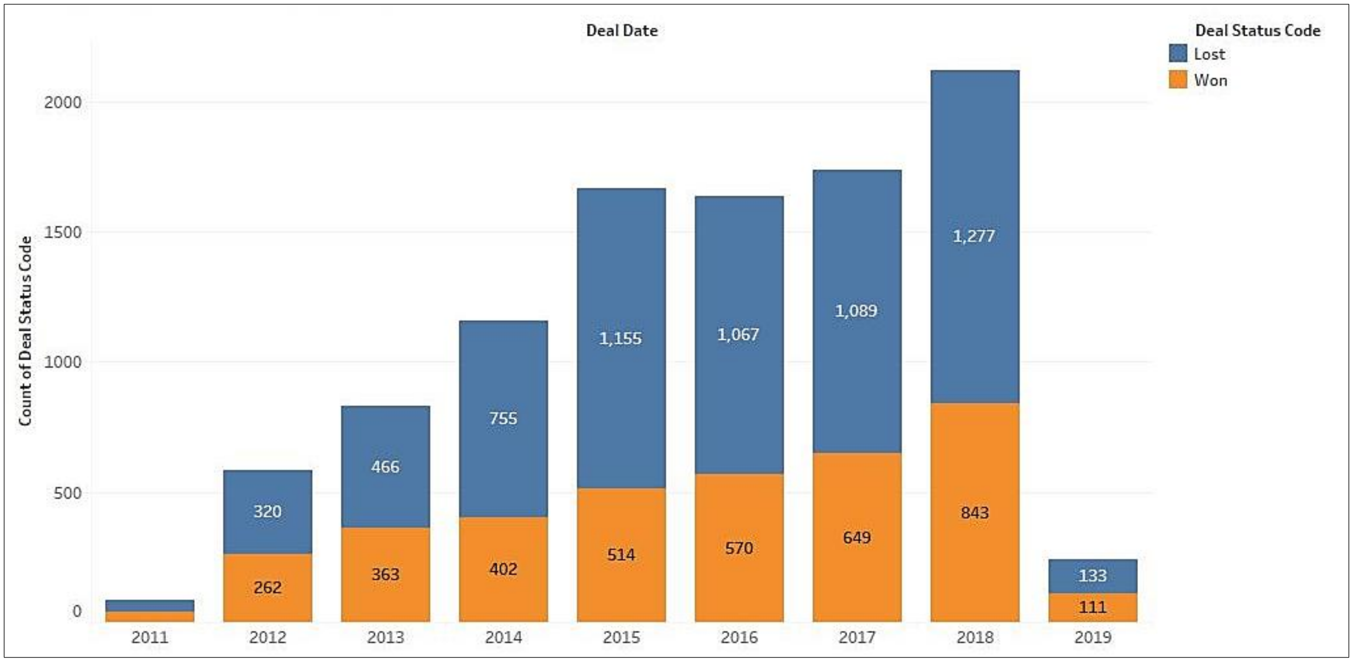
The above graph indicates the locations where the winning and losing bids were placed in Deal Cost and the number of bids respectively.

The Top 5 Bidding Locations:

Location	Won	Lost	Percentage of Wins
Location 10	1130	2230	33.63 %
Location 5	936	1392	40.20 %
Location 1	660	912	41.45 %
Others	391	426	47.86 %
Location 3	200	356	35.91 %

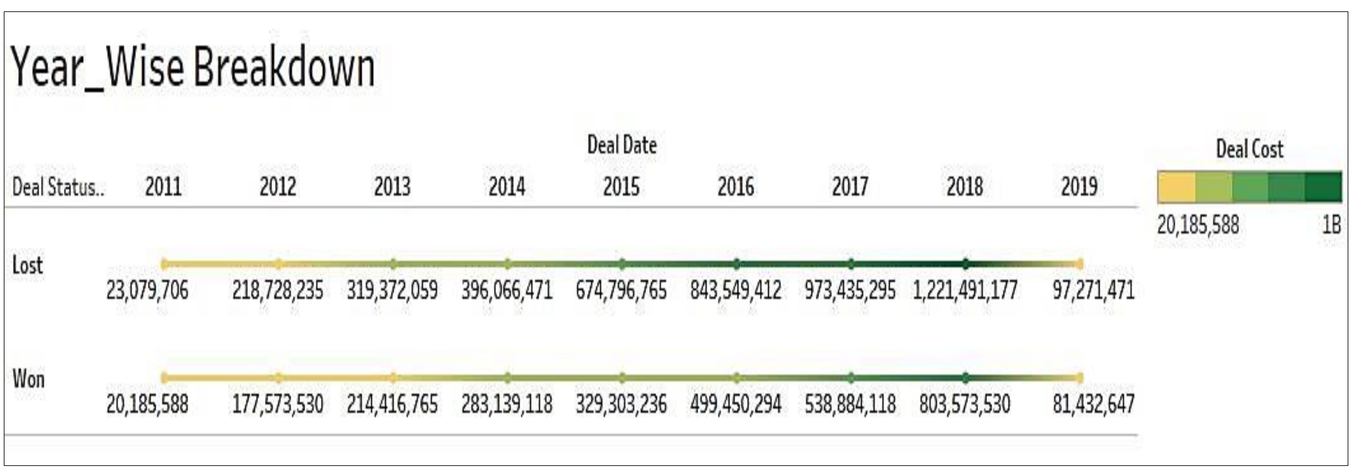
➢ Deal Date

The dataset contains bidding information from 2011 to mid-2019.



The graph distinctly makes it clear about the company's growth in bidding numbers over the years.

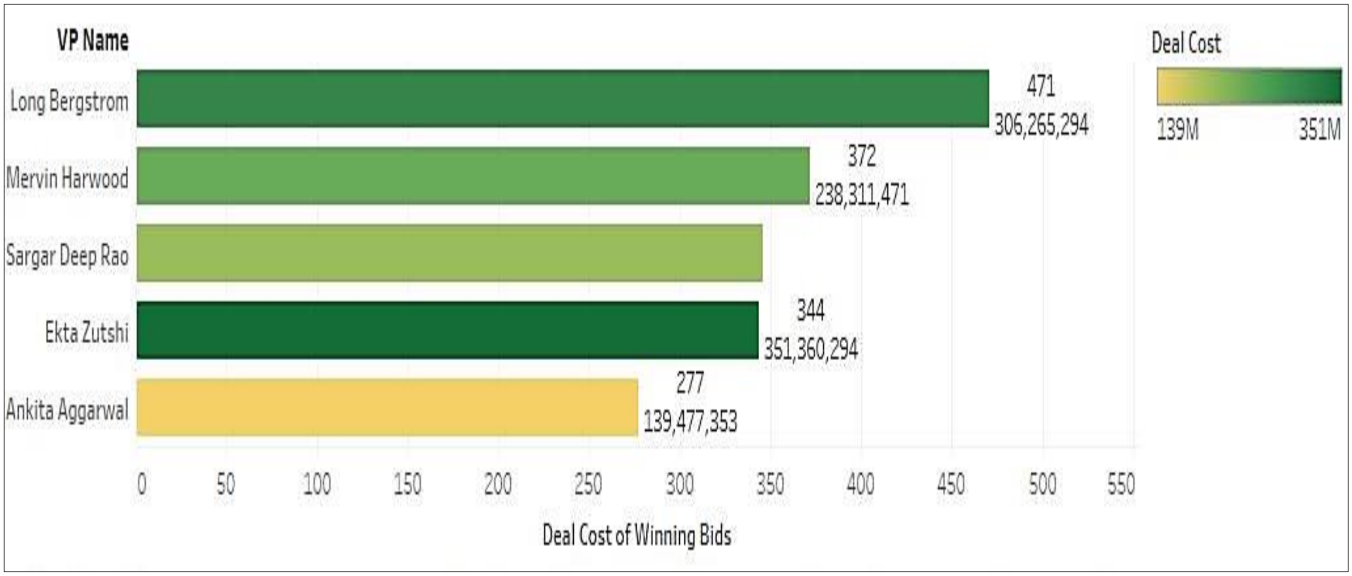
The year 2018 has been predominantly good in the company's overall performance.



The bids cost over the years is briefly explained in the above graph.

➢ VP / Senior Manger

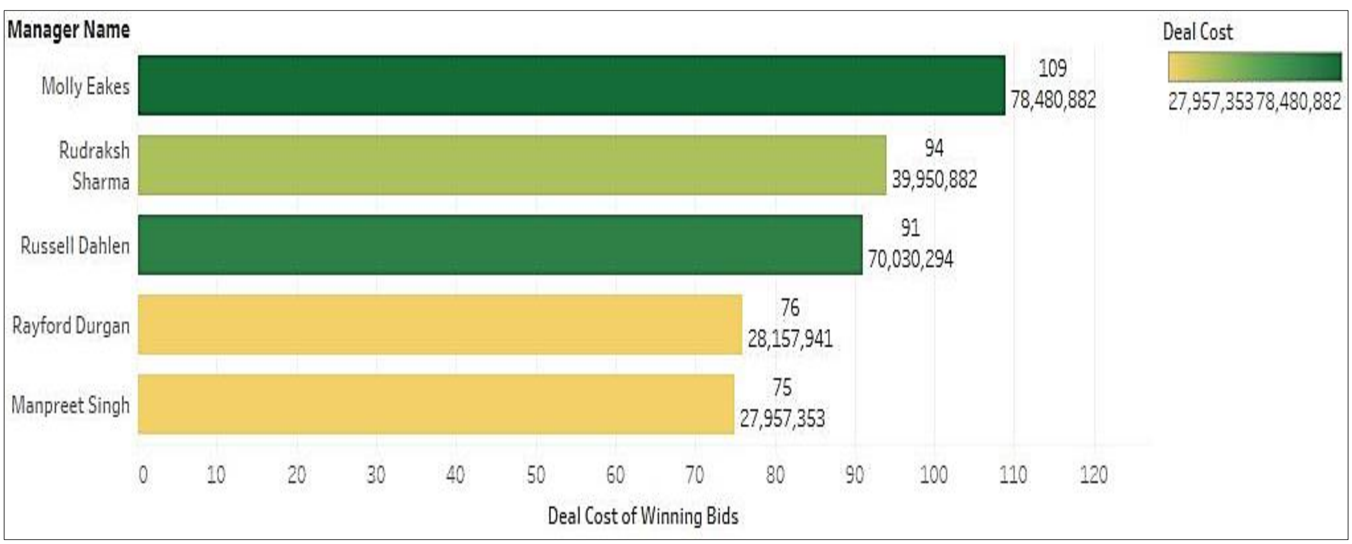
The company has around 43 VP/ Senior Managers for bidding partners.



The graph marks the Top 5 Bidding Senior Managers, with the total Deal cost bid by them over the years with the number of winning bids.

➢ Manager

The company has around 278 Managers for bidding partners.



The graph marks the Top 5 Bidding Managers, with the total Deal cost bid by them over the years with the number of winning bids.

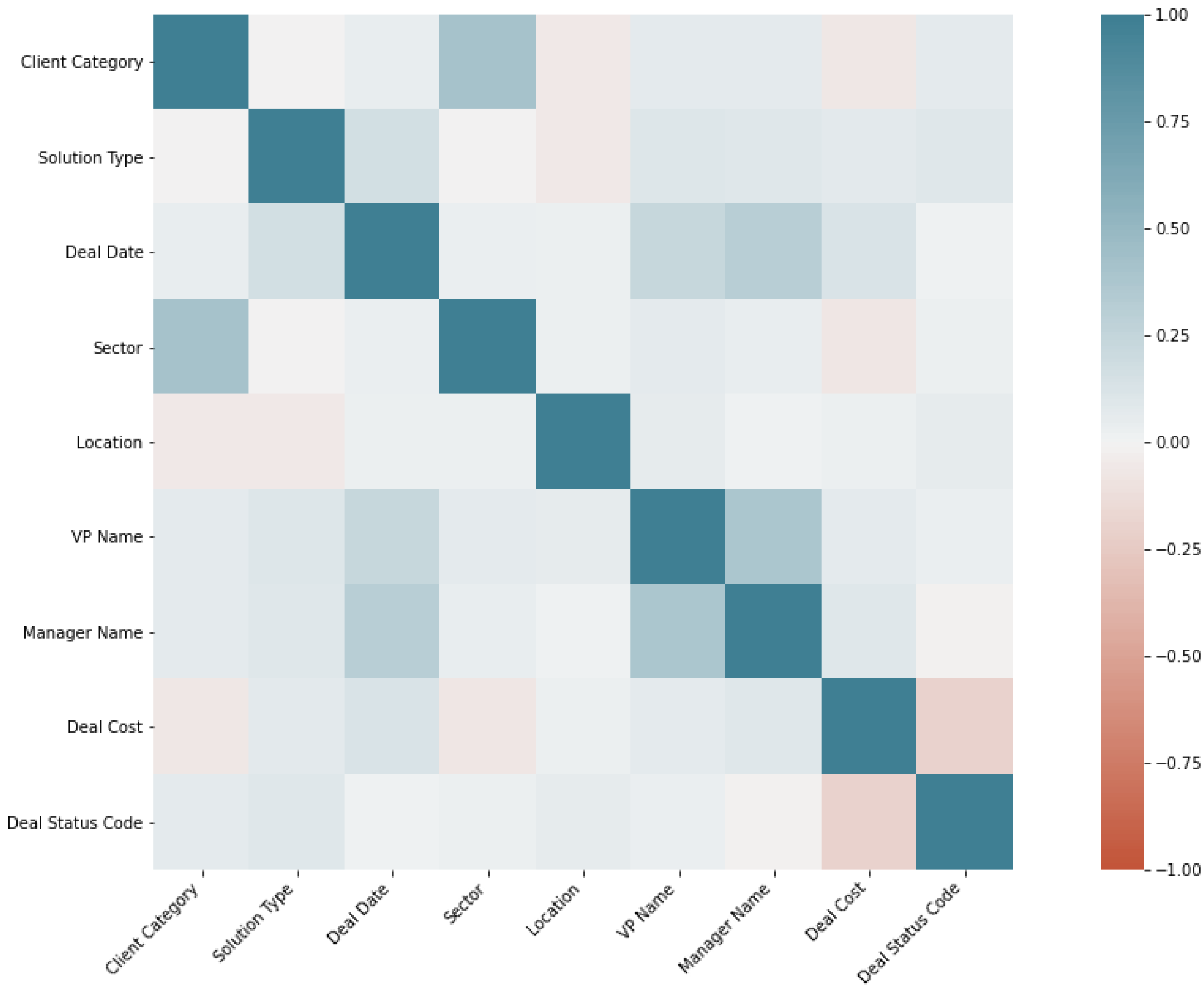
CORRPLOT

Since the given dataset mostly contains object type of data, to draw a corrplot

The dataset was factorized using Pearson method,

```
df = fullraw.apply( lambda x : pd.factorize(x)[0]).corr(method = "pearson", min_periods = 1)
```

- The Corrplot of the dataset reveals a good correlation among Sector and Client Category, and the Deal Cost and Deal Status Code to be negatively corelated.
- The Manager Name and Vp Name shows a higher rate of positive correlation.



5.DATA CLEANING

Data cleaning is the **process of identifying, deleting, and/or replacing inconsistent or incorrect information from the database**. This technique ensures high quality of processed data and minimizes the risk of wrong or inaccurate conclusions.

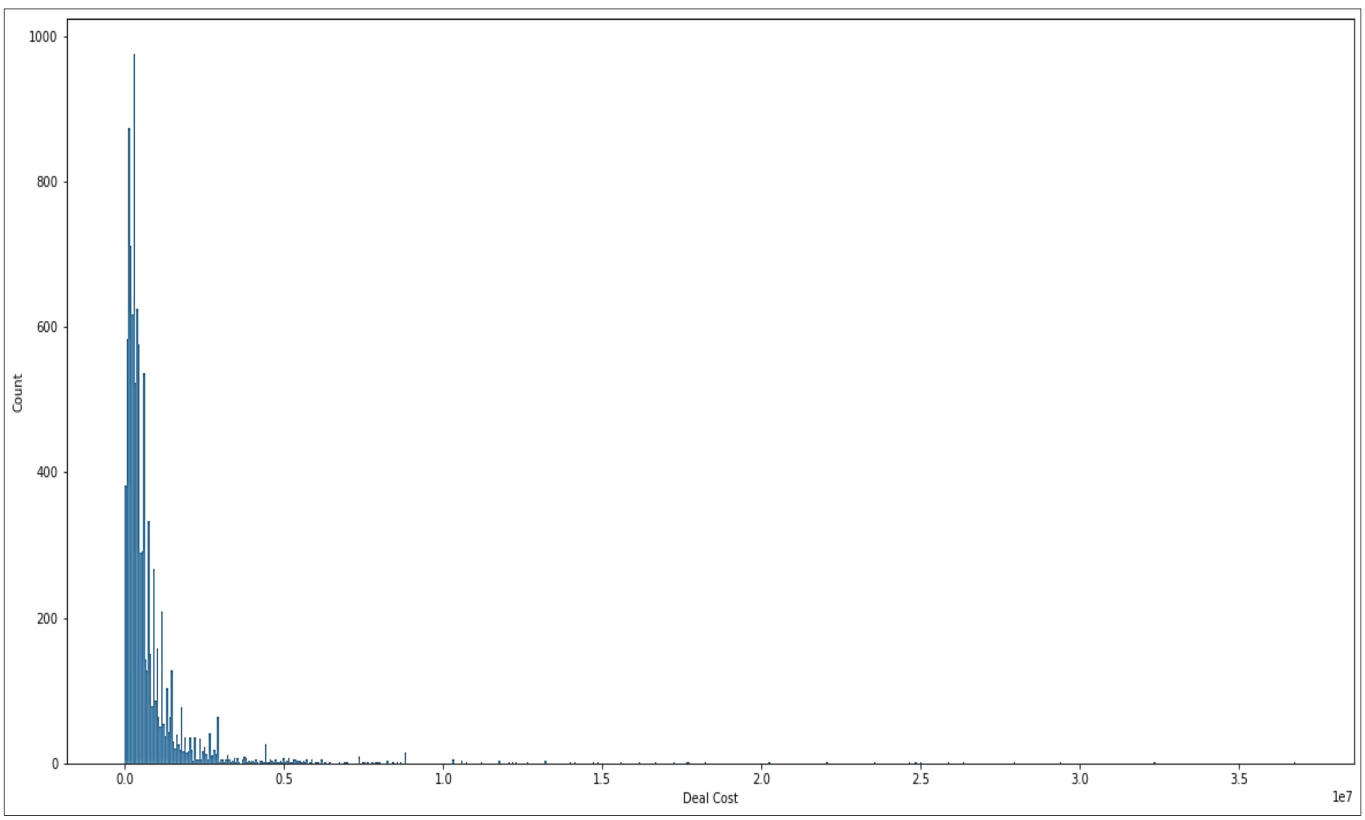
Handling Skewed data: **Deal Cost**

The Deal Cost column consisted of 246 Zero values, which was treated by replacing them with median value.

The median value is 3,82,352.9

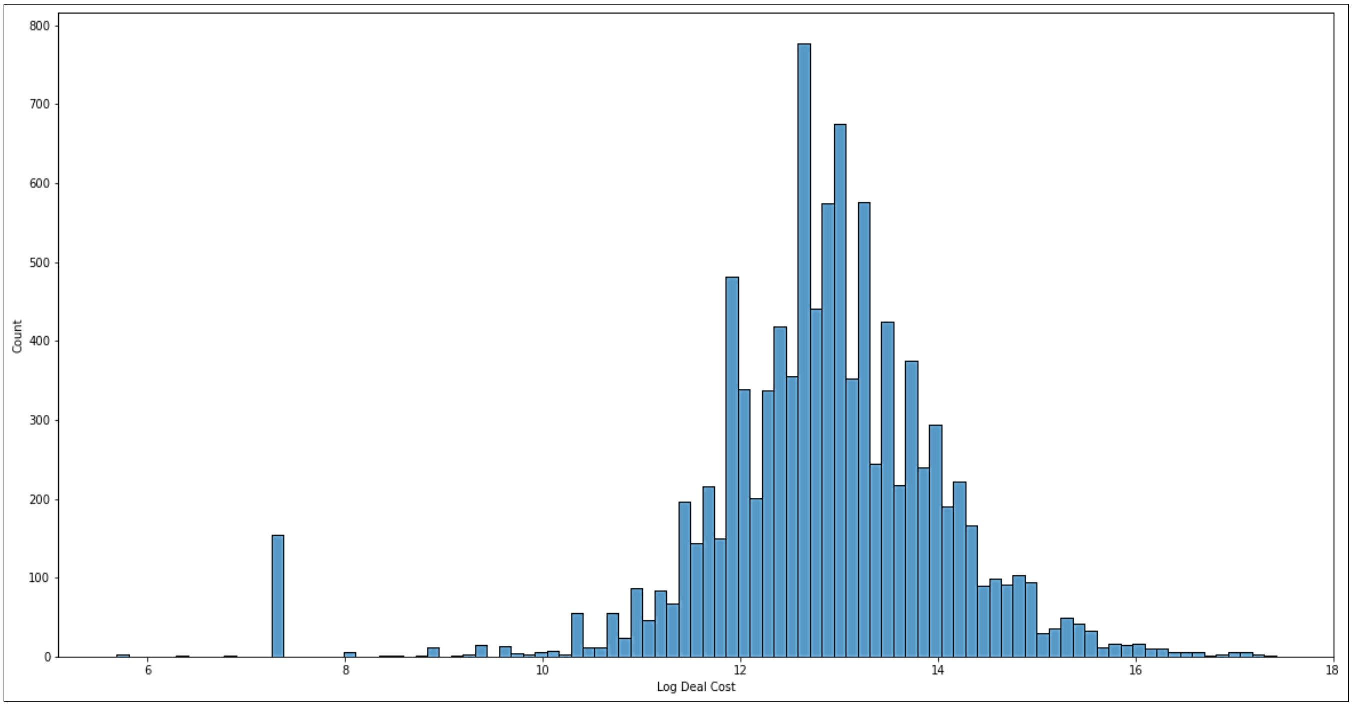
```
fullraw['Deal Cost'].replace(0.00, tempmedian, inplace = True)
```

The Deal Cost was Right Skewed.



Hence the Log Value of Deal Cost column was preferred for normalizing the data, and Deal Cost column dropped therein.

```
fullraw['Log Deal Cost'] = np.log(fullraw['Deal Cost'])
```



Thus, upon replacing Zero Values and considering the Log of Deal Cost, the data column now looks normalized.

Dropping Columns: **Deal Date, Deal Cost, VP Name, Manager Name**

Since the Target was to obtain the best bidding pairs of VP & Managers the respective columns were merged into single column and the previous ones removed.

```
fullraw['Vp_Manager'] = fullraw["VP Name"] + " " + fullraw["Manager Name"]
```

These columns were dropped from the dataset to reduce burden on the model.

```
fullraw = fullraw.drop(["Deal Cost"], axis = 1)

fullraw = fullraw.drop(['Deal Date'], axis = 1)

fullraw = fullraw.drop(["VP Name"], axis = 1)

fullraw = fullraw.drop(["Manager Name"], axis = 1)
```

Hence, After Data Cleaning and some Pre-processing the final data used for Model building consists of 10061 rows x 7 columns.

	Client Category	Solution Type	Sector	Location	Deal Status Code	Log Deal Cost	Vp_Manager
0	Telecom	Solution 7	Sector 24	L5	Won	11.918391	Ekta Zutshi Gopa Trilochana
1	Telecom	Solution 7	Sector 24	L5	Won	13.520745	Ekta Zutshi Gopa Trilochana
2	Internal	Solution 59	Sector 20	Others	Lost	11.002100	Ekta Zutshi Russell Dahlen
3	Internal	Solution 59	Sector 20	Others	Lost	11.002100	Ekta Zutshi Russell Dahlen
4	Internal	Solution 32	Sector 20	Others	Lost	11.300751	Ekta Zutshi Russell Dahlen
...	...	...	...	...	...	...	...
10056	Power ind	Solution 9	Sector 9	L5	Lost	13.284882	Rudraksh Sharma Rudraksh Sharma
10057	Internal	Solution 6	Sector 20	Others	Won	13.563271	Rudraksh Sharma Sharavan Singh
10058	Power ind	Solution 9	Sector 9	L5	Lost	13.284882	Rudraksh Sharma Rudraksh Sharma
10059	Power ind	Solution 62	Sector 9	L5	Won	14.928045	Man Suddeth Cleotilde Biron
10060	Others	Solution 9	Sector 12	L10	Lost	11.898588	Son Mcconnaughy Tarun Garg

10061 rows × 7 columns

6.MODEL BUILDING

The model building process involves **setting up ways of collecting data, understanding and paying attention to** what is important in the data to answer the questions you are asking, finding a statistical, mathematical or a simulation model to gain understanding and make predictions.

Since the given statement is a Classification problem, we chose the following algorithms for Building our model.

- **Logistic Regression:**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

- **Decision Tree:**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

- **Random Forest:**

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

- **XGBoost:**

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

DATA PRE-PROCESSING

RECODING DEPENDANT VARIABLE

The dependant Variable "Deal Status Code" predicts the Win or Lost output of a given bid, which is numerated for Model Building.

```
fullraw['Deal Status Code'] = np.where(fullraw["Deal Status Code"] == "Won", 1, 0)
```

INDEPENDANT AND TARGET VARIABLE SEPARATION

To perform Feature Engineering techniques on the dataset, it is divided into two separate data frames with one set containing the Independant Variables and the other with Target variable.

```
X = fullraw.drop("Deal Status Code", axis = 1);X
```

Client Category	Solution Type	Sector	Location	Log Deal Cost	Vp_Manager
Telecom	Solution 7	Sector 24	L5	11.918391	Ekta Zutshi Gopa Trilochana
Telecom	Solution 7	Sector 24	L5	13.520745	Ekta Zutshi Gopa Trilochana
Internal	Solution 59	Sector 20	Others	11.002100	Ekta Zutshi Russell Dahlen
Internal	Solution 59	Sector 20	Others	11.002100	Ekta Zutshi Russell Dahlen
Internal	Solution 32	Sector 20	Others	11.300751	Ekta Zutshi Russell Dahlen

```
y = fullraw["Deal Status Code"];y
```

```
1
1
0
0
0
```

FEATURE ENGINEERING

TARGET ENCODING- CATEGORICAL DATA

The Dataset in majority contains Categorical Data, thus to give out the best possible outcomes, the Ordinal data is numerated with the Mean value obtained in relation with the Target variable (‘Deal Status Code’) using the Target-Encoding Method, substituting for the Dummy Variable process.

```
cols = ['Client Category', 'Solution Type', 'Sector', 'Location', 'VP_Manager']
```

```
from sklearn.base import BaseEstimator, TransformerMixin

class TargetEncoder(BaseEstimator, TransformerMixin):

    def __init__(self, cols=None):

        if isinstance(cols, str):
            self.cols = [cols]
        else:
            self.cols = cols

    def fit(self, X, y):

        if self.cols is None:
            self.cols = [col for col in X
                         if str(X[col].dtype)=='object']

        for col in self.cols:
            if col not in X:
                raise ValueError('Column \''+col+'\' not in X')

        self.maps = dict()
        for col in self.cols:
            tmap = dict()
            uniques = X[col].unique()
            for unique in uniques:
                tmap[unique] = y[X[col]==unique].mean()
            self.maps[col] = tmap

        return self

    def transform(self, X, y=None):

        Xo = X.copy()
        for col, tmap in self.maps.items():
            vals = np.full(X.shape[0], np.nan)
            for val, mean_target in tmap.items():
                vals[X[col]==val] = mean_target
            Xo[col] = vals
        return Xo


    def fit_transform(self, X, y=None):

        return self.fit(X, y).transform(X, y)
```

Once the numerical substitutes are generated for the ordinal data, the model is fit on the dataset using **TargetEncoder**.

```
te   = TargetEncoder()
X_te = te.fit_transform(X, y)

X_te.sample(10)
```

Thus, the Final dataset for Model Building is ready and it is completely converted into numerical form.

Client Category	Solution Type	Sector	Location	Log Deal Cost	Vp_Manager
0.339434	0.322176	0.391014	0.359712	12.368592	0.400000
0.342826	0.301370	0.395349	0.402062	12.928207	0.363636
0.353963	0.282087	0.391014	0.367847	12.080910	0.372549
0.424242	0.370518	0.452575	0.402062	12.663125	0.666667
0.353963	0.480342	0.391014	0.336310	12.997200	0.688889
0.470270	0.320866	0.331137	0.336310	13.057354	0.248826
0.342826	0.564976	0.391014	0.419847	13.911890	0.680000
0.339434	0.262774	0.503234	0.336310	12.368592	0.360000
0.339434	0.564976	0.391014	0.367847	12.814879	1.000000
0.522696	0.222727	0.340961	0.478580	12.475201	1.000000

SAMPLING

The dataset was split into 70:30 as Train: Test respectively.

Train set has 7042 rows.

Test set has 3019 rows.

7.TESTING & VALIDATION

Each Model Creation process involved Prediction and Fit of the Model and generation of Confusion matrix of the same.

Some models were tuned based on **Hyper-Parametric Tuning Process**.

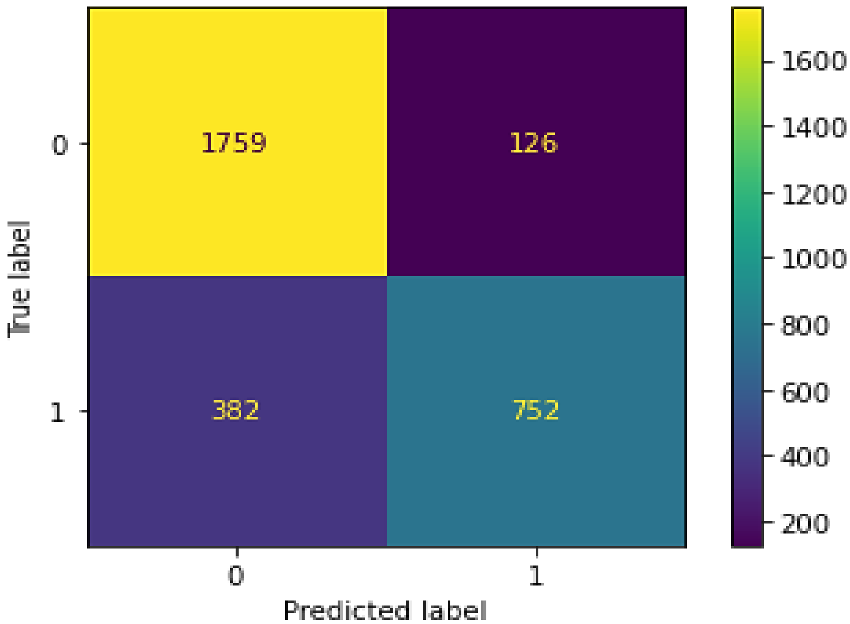
The result of the Final predictions are as follows:

Model Name	Precision	Recall	Accuracy	Roc-Auc	True-Loss
Logistic Regression	0.73	0.73	0.73	0.80	3.99×10^6
Decision Tree	0.82	0.82	0.82	0.99	2.67×10^6
Random Forest	0.83	0.83	0.83	0.99	2.65×10^6
XgBoost	0.81	0.81	0.81	0.97	2.68×10^6

(* The values stated for Precision and Recall are Weighted Average.)

From the above table it is clear that the **Random Forest** algorithm outperformed the rest of the models.

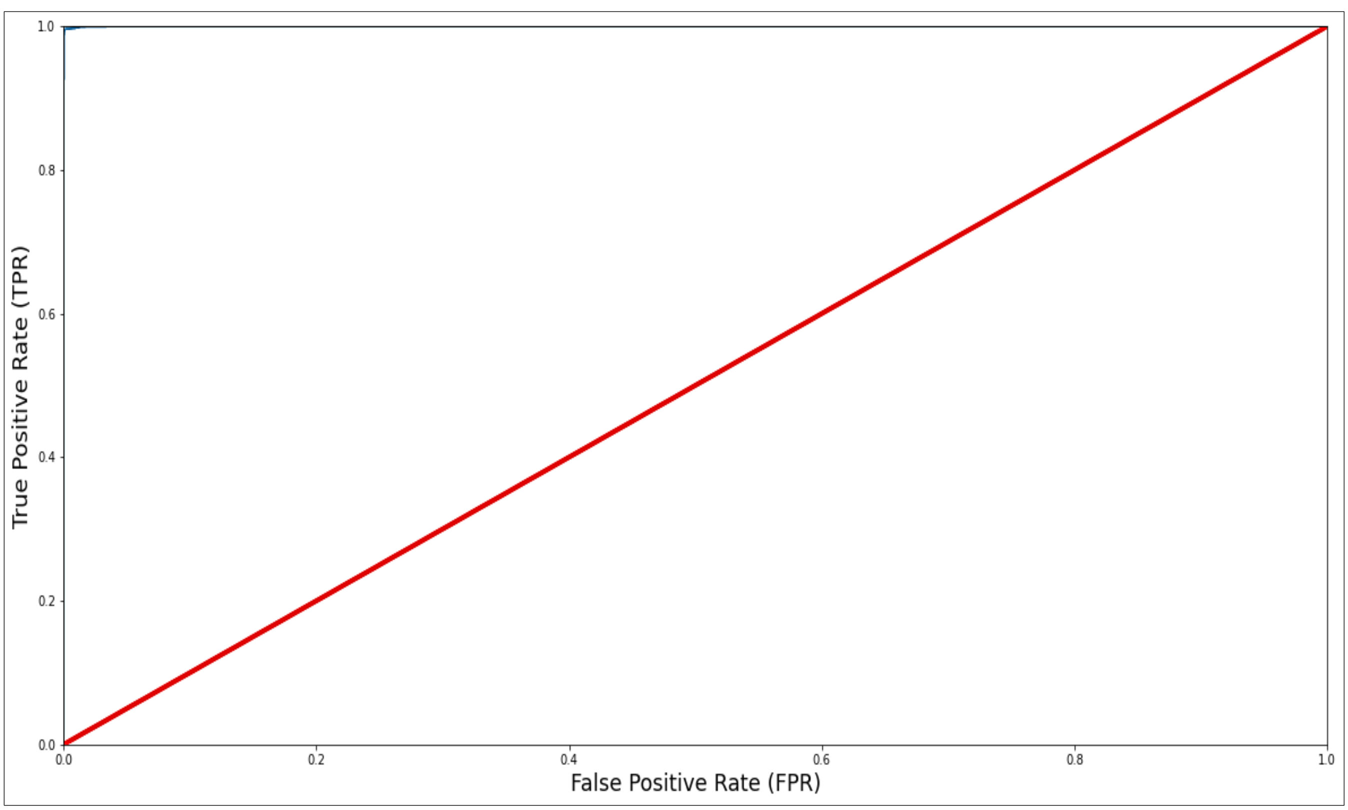
The Confusion matrix yielded for **Random Forest**:



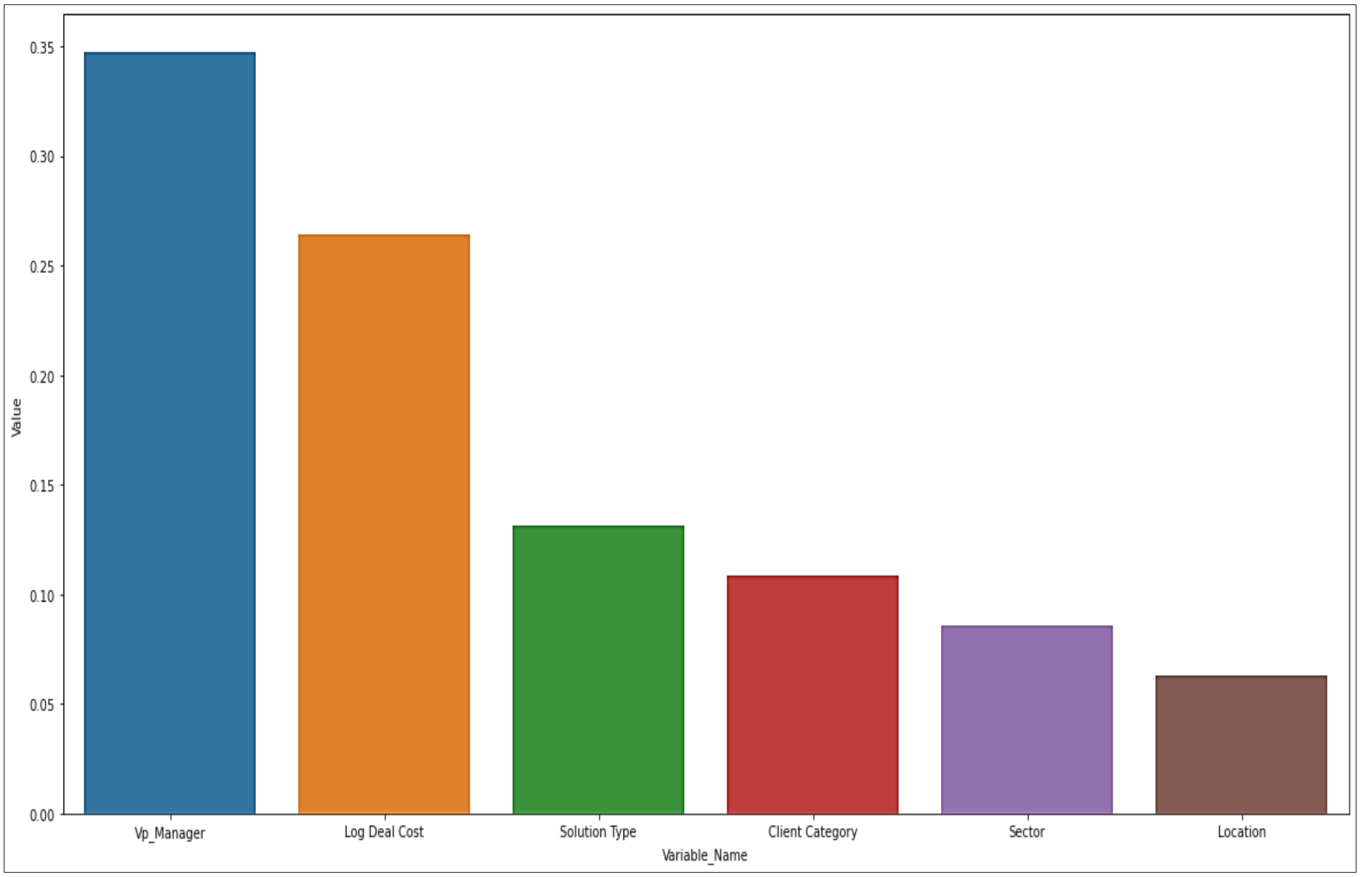
0's Predicted = 1759

1's Predicted = 752

The Roc_Auc Curve is scored at 99.9%



Visualization of the Important Variables:



The important variable dataset based on feature importance's was extracted into a new csv file.

	Value	Variable_Name
5	0.347190	Vp_Manager
4	0.263903	Log Deal Cost
1	0.131594	Solution Type
0	0.108464	Client Category
2	0.086113	Sector
3	0.062737	Location

8.RECOMMENDATIONS

The Top 5 Bidding VP and Manager Partnership

- Top 5 Recommendation of VP and Manager is based on the following:

  I) Win % (Total deals won by pair/ Total deals done by pair)

  II) Consistency (Total deals won by pair/ Total number of won deals)

  III) Efficiency (Win% x Consistency)

- Our recommendation is based on **Efficiency**, for unbiased selection over deals and win numbers and vice-versa.

VP Name	Manager	Total Deals	Total wins	Win %	Consistency	Efficiency
Long Bergstrom	Russell Dahlen	105	75	71.42	0.01997	1.4266
Ekta Zutshi	Neeraj Kumar	46	40	86.95	0.01065	0.9263
Neeraj Kumar	Vinay Kumar	75	51	68	0.01358	0.9235
Neeraj Kumar	Molly Eakes	144	62	43.05	0.01651	0.7109
Rahul Bajpai	Rudraksh Sharma	198	72	36.36	0.01917	0.6972

9.INSIGHTS INTO DATA

Suggestions for taking better Decisions.

- The Deals bid by the Company are in majority under 2 MUSD (approx. 9000 bids).

- Hence, we recommend **Top 3 Performing** Senior Manager's and Manager's bidding under 2Mil.

Senior Manager's:

- ✓ Long Bergstrom
- ✓ Ekta Zutshi
- ✓ Sagardeep Rao

Manager's:

- ✓ Rayford Durgan
- ✓ Molly Eakes
- ✓ Rudraksh Sharma

10.CONCLUSION

- Corporate project bids are a valuable part for functioning of the business.
- Our ML model assists in predicting the win or loss of a bid, for a potential client.
- From our Analysis, the VP and Manager partnership form a significant influence over the probability of deal to win/lose, followed by Deal Cost and the rest of the variables forming minimal significance.
- Thus, it is imperative for the Organization to staff VP and Manager's with **higher efficiency** to deals which pose higher significance.

FUTURE WORK

The Organization can **capture the feedback** behind a Win/Loss of a bid, thus enabling us to analyze the reasons for a bid's success/ failure, and thereby helping the organization enhance its chances.

REFERENCES

- ROC Curve: Making way for correct diagnosis
- Better Heatmaps and Correlation Matrix Plots in Python
- XGBoost Documentation — xgboost 1.5.1 documentation
- https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/
- https://brendanhasz.github.io/2019/03/04/target-encoding
- https://www.geeksforgeeks.org/feature-encoding-techniques-machine-learning/
- Prescriptive analytics: An insider's guide