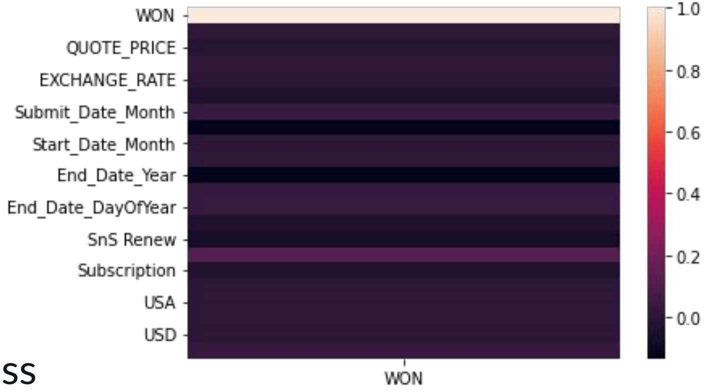
IBM_CAO_Data_Science_Challenge

Palak Bansal, MS in Data Science, NYU

Data

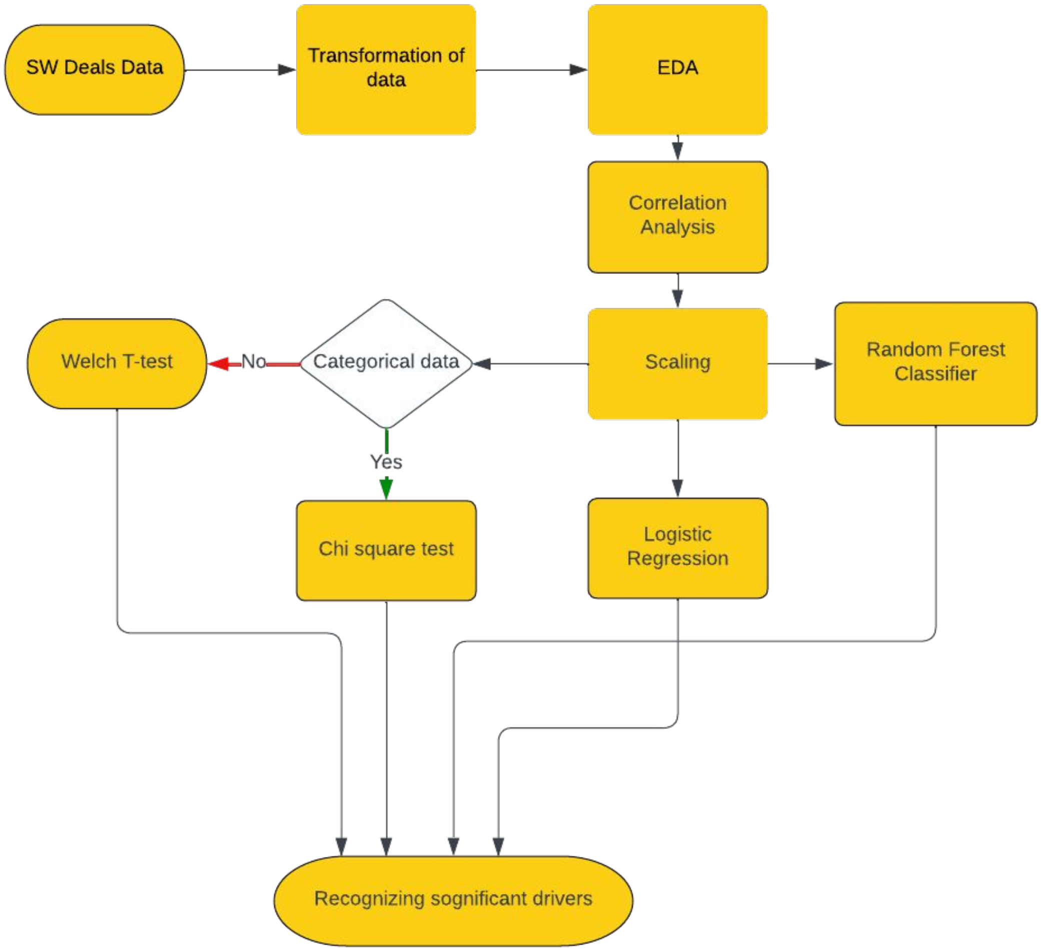


SW Deals Data:

- ✓ No missing data
- ✓ Classes balanced for win and loss
- ● 5 categorical features, 5 continuous features, 3 datetimes
- ● Correlation values aren't strong

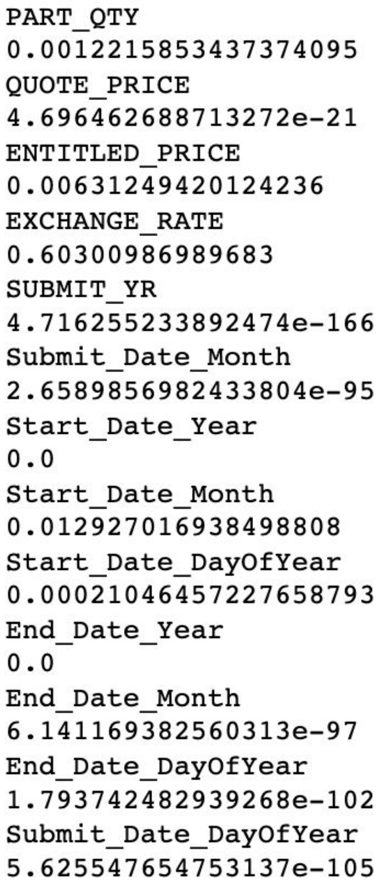
Competitors/Comments Data:

- ✓ All comment types are company names
- ● Multiple company names in the comments are separated differently.
- ● Only 21264 from SW deals have corresponding comments


SW Deals Data
Transformation of data
EDA
Correlation Analysis
Scaling
Random Forest Classifier
Categorical data
No
Welch T-test
Yes
Chi square test
Logistic Regression
Recognizing sognificant drivers

Win/Loss drivers

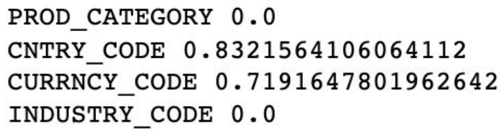
Significance Tests



```
PART_QTY
0.0012215853437374095
QUOTE_PRICE
4.696462688713272e-21
ENTITLED_PRICE
0.00631249420124236
EXCHANGE_RATE
0.60300986989683
SUBMIT_YR
4.716255233892474e-166
Submit_Date_Month
2.6589856982433804e-95
Start_Date_Year
0.0
Start_Date_Month
0.012927016938498808
Start_Date_DayOfYear
0.00021046457227658793
End_Date_Year
0.0
End_Date_Month
6.141169382560313e-97
End_Date_DayOfYear
1.793742482939268e-102
Submit_Date_DayOfYear
5.625547654753137e-105
```

Results from Welch t-test (checked for different variances) performed on win/loss. Analysing the p-values, the significant features in decreasing order are:

- Start Date year
- End Date year
- Submit Year
- Submit Date's day of year
- End date's day of the year
- End date month
- Submit Date Month
- Quote Price
- Start Day's day of year

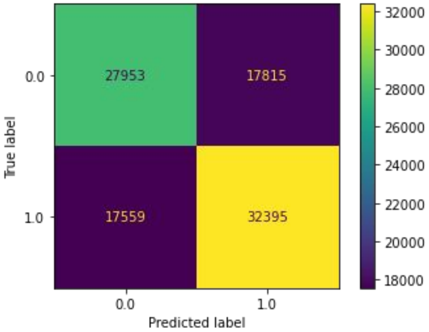


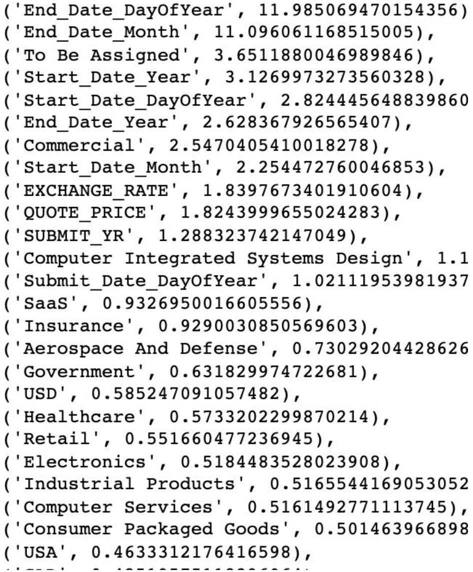
```
PROD_CATEGORY 0.0
CNTRY_CODE 0.8321564106064112
CURRNCY_CODE 0.7191647801962642
INDUSTRY_CODE 0.0
```

Results from chi-square test (for categorical variables) performed on win/loss. Analysing the p-values, the significant features in decreasing order are:

- Product Category
- Industry Code

Logistic Regression

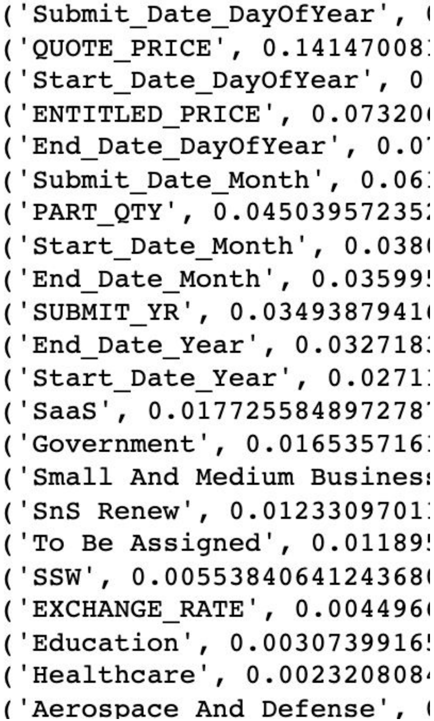




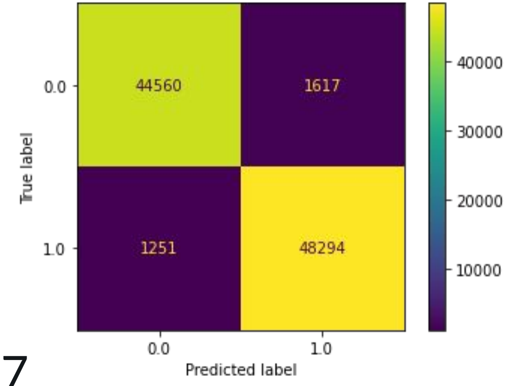
```
('End_Date_DayOfYear', 11.985069470154356)
('End_Date_Month', 11.096061168515005),
('To Be Assigned', 3.6511880046989846),
('Start_Date_Year', 3.1269973273560328),
('Start_Date_DayOfYear', 2.824445648839860
('End_Date_Year', 2.628367926565407),
('Commercial', 2.5470405410018278),
('Start_Date_Month', 2.254472760046853),
('EXCHANGE_RATE', 1.8397673401910604),
('QUOTE_PRICE', 1.8243999655024283),
('SUBMIT_YR', 1.288323742147049),
('Computer Integrated Systems Design', 1.1
('Submit_Date_DayOfYear', 1.02111953981937
('SaaS', 0.932695001660556),
('Insurance', 0.9290030850569603),
('Aerospace And Defense', 0.7302920442862б
('Government', 0.631829974722681),
('USD', 0.585247091057482),
('Healthcare', 0.5733202299870214),
('Retail', 0.551660477236945),
('Electronics', 0.5184483528023908),
('Industrial Products', 0.5165544169053052
('Computer Services', 0.5161492771113745),
('Consumer Packaged Goods', 0.501463966898
('USA', 0.4633312176416598),
```

- Area Under Curve = 0.63
- Almost equal training and testing metrics -> Model is underfitting.
- The resulting coefficients would give feature importance.
- Results mostly consistent with the values from the t-tests.
- Decreasing order of absolute value of coefficients in the picture, giving the significant drivers in decreasing order.
- Dates are consistently a significant driver
- Better model:
  -> Using a more complicated model.
  -> More relevant features about competitors and SW Deals.
  -> Identifying and handling outliers(noise).

Random Forest Classifier



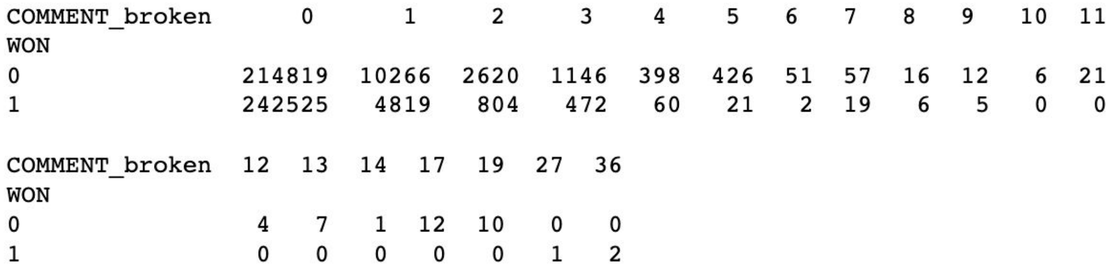
```
('Submit_Date_DayOfYear', 
('QUOTE_PRICE', 0.14147008
('Start_Date_DayOfYear', 0
('ENTITLED_PRICE', 0.07320
('End_Date_DayOfYear', 0.0
('Submit_Date_Month', 0.06
('PART_QTY', 0.04503957235
('Start_Date_Month', 0.038
('End_Date_Month', 0.03599
('SUBMIT_YR', 0.0349387941
('End_Date_Year', 0.032718
('Start_Date_Year', 0.0271
('SaaS', 0.017725584897278
('Government', 0.016535716
('Small And Medium Busines
('SnS Renew', 0.0123309701
('To Be Assigned', 0.01189
('SSW', 0.0055384064124368
('EXCHANGE_RATE', 0.004496
('Education', 0.0030739916
('Healthcare', 0.002320808
('Aerospace And Defense', 
```



- Training score: 0.9683926808501747
- Testing score: 0.9682027412537315
- Dates are among the significant drivers again, Random Forest creates a much more accurate model -> feature importances more reliable.
- Quote Price is a strong driver.
- Entitled price is an important feature for the model, but also highly correlated with quote price.

Combining with comments data

- Features extracted -> Competitors count
- Improves results of Logistic and Random Forest classifier, but marginally.
- Assumption: Deals having no comments have 0 competitors.
- Possible reasoning: Competitor data present for very few deals, competitor count is mostly 0.
- As the competition increases, SWHub tends to lose more deals as shown in picture.
- Additional data :
  -> Competitor information on more deals.
  -> Competitor's quote prices and dates.



```
COMMENT_broken       0      1     2     3    4    5   6   7   8   9   10  11
WON
0               214819  10266  2620  1146  398  426  51  57  16  12    6  21
1               242525   4819   804   472   60   21   2  19   6   5    0   0

COMMENT_broken  12  13  14  17  19  27  36
WON
0                4   7   1  12  10   0   0
1                0   0   0   0   0   1   2
```