IBM chief analytics office Data Challenge

Shrey Arora

Biggest Revenue Contributor in 2021

Product	Revenue
SSW	9.66E+08
SaaS	1.61E+09
SnS Renew	5.28E+09
Subscription	1.11E+00

SnS Renew contributed the most to the revenue in 2021

```
data_2021 = data[data["SUBMIT_YR"]==2021]
revenues = pd.DataFrame()
for n,g in data_2021.groupby("PROD_CATEGORY",0):
    revenues = revenues.append({"Product":n,"Revenue": g["QUOTE_PRICE"].sum()},ignore_index = True)

revenues.iloc[revenues["Revenue"].argmax()]
```

Most significant driver’s for win/loss

Correlational Analysis

- Test performed: Chi square and ANOVA

  Conclusion - Exchange Rate, CNTRY_CODE, CURRNCY CODE are least significant on win/loss status

ML Results

- Random Forest gave 'PART_QTY', 'QUOTE_PRICE', 'ENTITLED_PRICE', 'SUBMIT_DATE', 'Duration' as the most significant variables.

  It predicted the sales with **72.3** percent accuracy

*I created a variable called Duration (Duration in days between start_date and end_date)

Performance Improvement

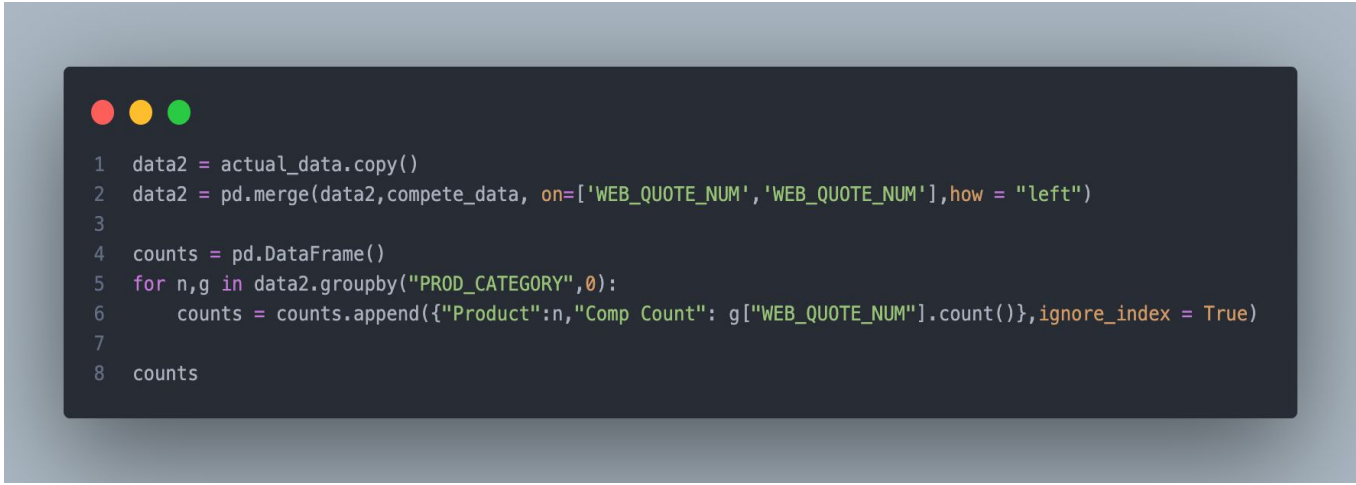
- Implement different regularization methods like LASSO, Ridge, elastic net with logistic regression.
- Grid Search for regularization parameters and model tuning parameters.
- Use ensemble methods to improve the accuracy of the model with a trade off in interpretability.
- Feature Engineering for e.g reduce number of classes in PART_NUM etc.

Additional Data to enhance the model

- Social/Political situation at customer location - As seen in covid 19 pandemic
- Market Trends/ Economic Conditions - People spend less during recession.
- Marketing/Sales Channel - Different marketing mediums have different conversion rates and can affect the prediction model.

Product Category with most number of average competitors

Product	Comp Count
SSW	50327
SaaS	99419
SnS Renew	323076
Subscription	5786



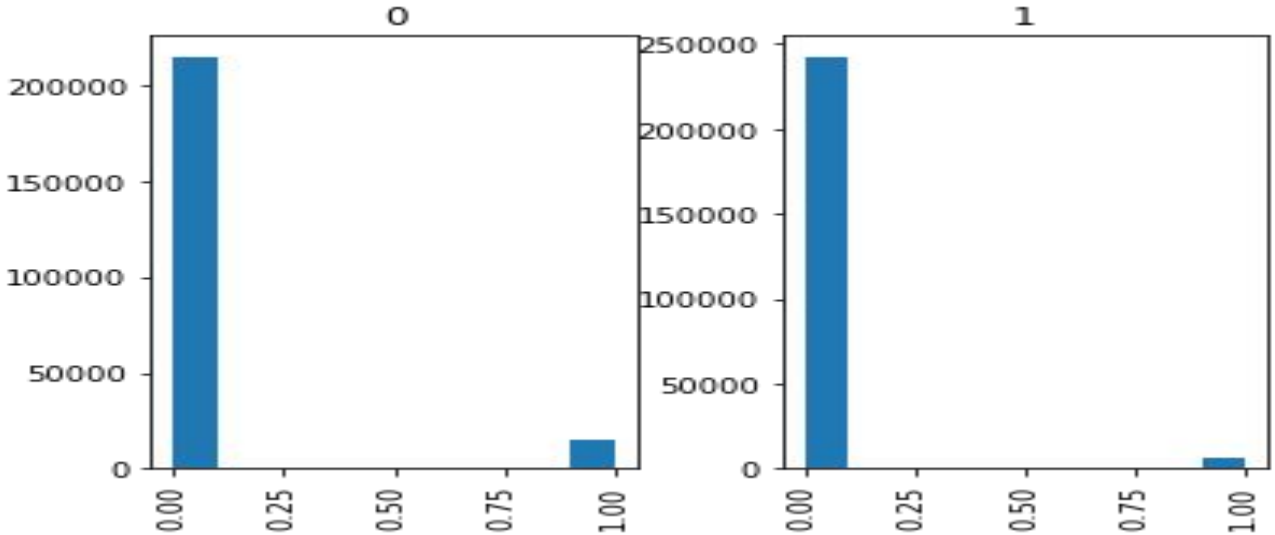
```
data2 = actual_data.copy()
data2 = pd.merge(data2,compete_data, on=['WEB_QUOTE_NUM','WEB_QUOTE_NUM'],how = "left")

counts = pd.DataFrame()
for n,g in data2.groupby("PROD_CATEGORY",0):
    counts = counts.append({"Product":n,"Comp Count": g["WEB_QUOTE_NUM"].count()},ignore_index = True)

counts
```

Sns Renew has most number of average competitors

*I assumed in Tabel2-SW Deals Comments each WEB_QUOTE_NUM and has average of one competitor per row.

There were missing data when I joined the two tables

Effect of Competitor on Win/Loss



The above plot represents count of Deals won and lost with competitors(Right) and without competitors (Left).