Technical Assessment

Data Scientist - 007234

Table of Contents

1. Tools/ Libraries Used……………………………………………………….Page 3
2. Problem Statement………………………………………………………….Page 5
3. Glossary…………………………………………………………………….Page 6
4. Analysis Task I……………………………………………………………...Page 7
    4.1. Approach towards the solution……………………………Page 5
    4.2. Metric Selection……………………………………………Page 5
    4.3. Results/ Insights……………………………………………Page 7
5. Analysis Task II……………………………………………………………..Page 10
6. Approach towards the solution……………………………Page 10
    6.1. Feature Engineering……………………………………….Page 10
    6.2. Algorithm Selection……………………………………….Page 11
    6.3. Performance Metric………………………………………..Page 11
    6.4. Results/ Insights……………………………………………Page 16
7. Github Repository.…………………………………………………………...Page 16

TOOLS/ LIBRARIES USED

Tool/ Technique	Used
Programming Language	Python
IDE	Jupyter Notebook (Open-Source)
Python Libraries	Pandas, Numpy, Statsmodels, scipy, warnings
Python data visualisation libraries	Matplotlib, pyplot, seaborn
Types of charts	Line Chart, scatter plots
Code Versioning	Github
Github URL	https://github.com/sumakshi/Bank-of-England.git

PROBLEM STATEMENT

A Supervision Manager has asked you to help in allocating scarce resources, and identify which firms their team should focus on. Supervisory resource may be allocated according to the following characteristics:

1. Firm size (i.e. the biggest firms need more attention)
2. Changing business profile (are firms' data changing substantially year-on-year?)
   Outliers from the norm (when looking at a single reporting period, does a firm deviate significantly from the average?)
3. Some typical metrics have been provided in the attached data sheets. These include:

   - Gross Written Premium (GWP) – total revenue written by an insurer. Equivalent of turnover for a non-insurance firm.
   - Net Written Premium (NWP) – GWP less reinsurance. NWP / GWP will show how much of the firm's risk is being passed on to reinsurers.
   - SCR coverage ratio – a measure of whether a firm is meeting its prudential capital requirements. Greater than 100% means the firm is holding enough capital to meet the requirement. The size of the buffer (i.e. surplus over 100%) can be important.
   - Gross claims incurred – a large cost to an insurer. Monitoring how these change over time for a firm is vital.
   - Net combined ratio – (incurred losses plus expenses) / earned premiums. This is a ratio that can indicate the profitability of a firm. If this is less than 100% it indicates a profit.

GLOSSARY

Terms	Abbreviation	Meaning
GWP	Gross Written Premium	Total turnover of the company
NWP	Net Written Premium	Total premium on insurance underwritten by an insurer during a specified period after the deduction of premium applicable to reinsurance.
SCR	Solvent Capital Ratio	EOF / SCR
EOF_SCR	Eligible Own Fund SCR	The SCR ratio is calculated using Eligible Own Funds most of the time.
EOAOL	Assets Over Liabilities	If a firm has more assets than liabilities then they are solvent.
NWP/GWP	NWP/GWP ratio	Insuring the insurers

TASK I

Using the data provided, please analyse this data using R or Python and produce a short report, including tables and charts, to highlight which firms should receive the most attention, according to the metrics above.

Approach:

Metrics Used:

NWP_GWP_AVG(Average of NWP/GWP ratio), SCR_ratio_AVG(Average of EOF/SCR ratio)

```
In [7]: #Calculating average for NWP/GWP ratio, GWP ratio, EOAOL, Total Liabilities, EOF_SCR and SCR
df2['NWP_GWP_AVG']=(df2['NWP_GWP_2016']+df2['NWP_GWP_2017']+df2['NWP_GWP_2018']+df2['NWP_GWP_2019']+df2['NWP_GWP_202
df2['GWP_AVG']=(df2['GWP_2016']+df2['GWP_2017']+df2['GWP_2018']+df2['GWP_2019']+df2['GWP_2020'])/5
df2['EOAOL_AVG']=(df2['EOAOL_2016']+df2['EOAOL_2017']+df2['EOAOL_2018']+df2['EOAOL_2019']+df2['EOAOL_2020'])/5
df2['Total_liabilities_AVG']=(df2['Total_liabilities_2016']+df2['Total_liabilities_2017']+df2['Total_liabilities_201
df2['EOF_SCR_AVG']=(df2['EOF_SCR_2016']+df2['EOF_SCR_2017']+df2['EOF_SCR_2018']+df2['EOF_SCR_2019']+df2['EOF_SCR_202
df2['SCR_AVG']=(df2['SCR_2016']+df2['SCR_2017']+df2['SCR_2018']+df2['SCR_2019']+df2['SCR_2020'])/5
```

```
In [8]: #Calculating standard deviation of NWP/GWP ratio for each year
df2['NWP_GWP_2016_AVG']=df2['NWP_2016'].sum()/len(df2['NWP_2016'])
df2['NWP_GWP_2016_SD'] = df2['NWP_GWP_2016']- df2['NWP_GWP_2016_AVG']
df2['NWP_GWP_2017_AVG']=df2['NWP_2017'].sum()/len(df2['NWP_2017'])
df2['NWP_GWP_2017_SD'] = df2['NWP_GWP_2017']- df2['NWP_GWP_2017_AVG']
df2['NWP_GWP_2018_AVG']=df2['NWP_2018'].sum()/len(df2['NWP_2018'])
df2['NWP_GWP_2018_SD'] = df2['NWP_GWP_2018']- df2['NWP_GWP_2018_AVG']
df2['NWP_GWP_2019_AVG']=df2['NWP_2019'].sum()/len(df2['NWP_2019'])
df2['NWP_GWP_2019_SD'] = df2['NWP_GWP_2019']- df2['NWP_GWP_2019_AVG']
df2['NWP_GWP_2020_AVG']=df2['NWP_2020'].sum()/len(df2['NWP_2020'])
df2['NWP_GWP_2020_SD'] = df2['NWP_GWP_2020']- df2['NWP_GWP_2020_AVG']
```

```
In [9]: #Calculating SCR ratio for each year
df2['SCR_ratio_2016']=df2['EOF_SCR_2016']/df2['SCR_2016']
df2['SCR_ratio_2017']=df2['EOF_SCR_2017']/df2['SCR_2017']
df2['SCR_ratio_2018']=df2['EOF_SCR_2018']/df2['SCR_2018']
df2['SCR_ratio_2019']=df2['EOF_SCR_2019']/df2['SCR_2019']
df2['SCR_ratio_2020']=df2['EOF_SCR_2020']/df2['SCR_2020']
```

Data Wrangling:

- Excluding firms with no business in the past 5 years

```
In [4]: #Removing Outliers(i.e. elcluding firms with no business in the past 5 years)
df2=df1[(df1['GWP_2016']>0) & (df1['GWP_2017']>0) & (df1['GWP_2018']>0) & (df1['GWP_2019']>0)& (df1['GWP_2020']>0)]
```

- Replaced Null values with 0
- Calculating average and standard deviation of values of the metrics provided

Results/ Insights:

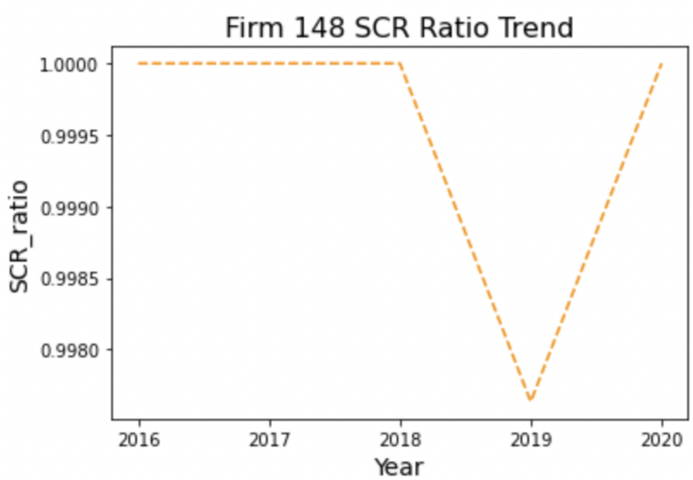
1. **Top 3 firms to focus immediately:**
   Firm 148, Firm 213 and Firm 183 respectively (Refer Cell 14) needs help and should be focused immediately as otherwise these would become insolvent because their average of SCR ratio is less than 1. The order for help is decided using their corresponding GWP value, i.e. the firm with high GWP value is more valuable and should be given priority.

Top 3 firms which require help immediately

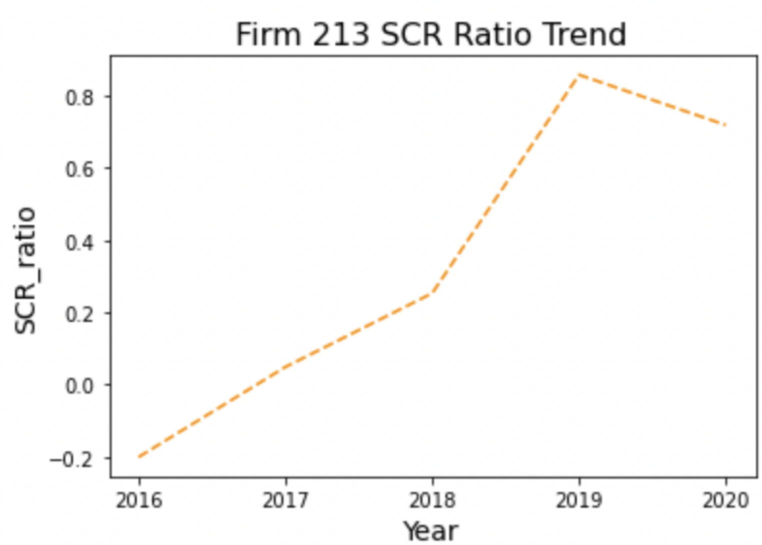
In [14]:
```
# Selecting top 3 firms having SCR ratio less than 1 i.e. firms which are about to become insolvent.
# Priority order is the GWP i.e.company with more turnover should be given priority
immediate=df5[df5['SCR_ratio_AVG']<1]
immediate.sort_values(by=['GWP_AVG'], ascending=False).head(3)
```

Out[14]:

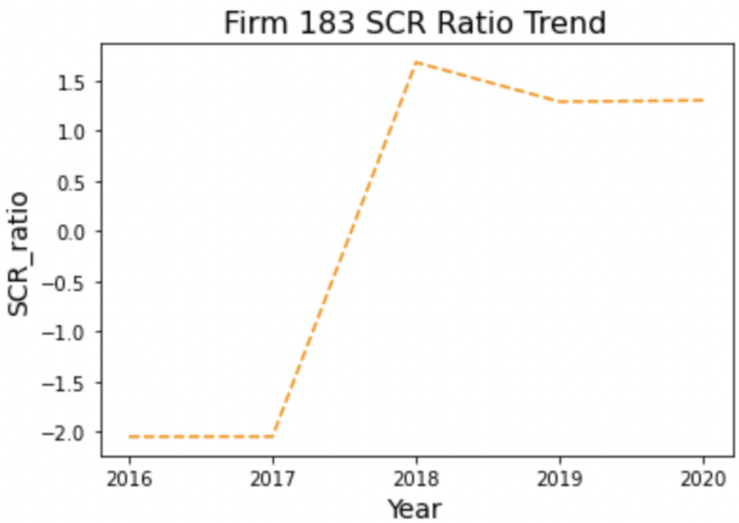
	Firm_Name	GWP_AVG	NWP_GWP_AVG	SCR_ratio_AVG	EOAOL_AVG	NWP_GWP_AVG_STD	SCR_ratio_AVG_STD
147	Firm 148	60.580624	0.882305	0.999527	182.178579	2.537884	1.535742
212	Firm 213	19.069063	1.000000	0.335523	89.751959	2.537884	1.535742
182	Firm 183	0.002746	1.000000	0.035835	16.890128	2.537884	1.535742



2. In 2019, Firm 148 had more liabilities than assets, i.e. their SCR ratio dropped suddenly. However, the ratio improved significantly in 2020 but there could be a random fall again. Since this firm has a GWP average of 60.580624(Highest), it becomes the most important firm.



3. For Firm 213, there was a sharp increase in the SCR ratio for a short period between 2018 and 2019. However, the SCR ratio is decreasing since 2019. Since this firm has a GWP average of 19.069063(Second highest), it becomes the second most important firm.



4. For Firm 183, there was a steep rise in the SCR ratio for a short period between 2017 and 2018. However, the ratio is decreasing since 2018. Since this firm has a GWP average of 0.002746, it becomes the least important firm out of 3 but significantly important when compared to the total of 325 firms.

Top 3 firms to focus in a short-term:

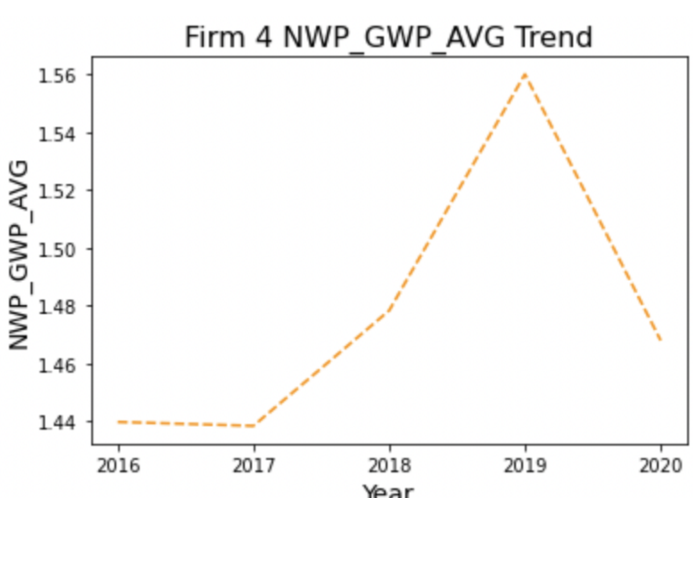
1. Firm 4, Firm 105, Firm 311 respectively (Refer Cell 18) needs help and should be focused in a short term as these firms are not completely/ properly insured because their average NWP/ GWP ratio is less than 1. The order for help is decided using their corresponding GWP value, i.e. the firm with high GWP value is more valuable and should be given priority.

Top 3 firms which require help in short time ¶

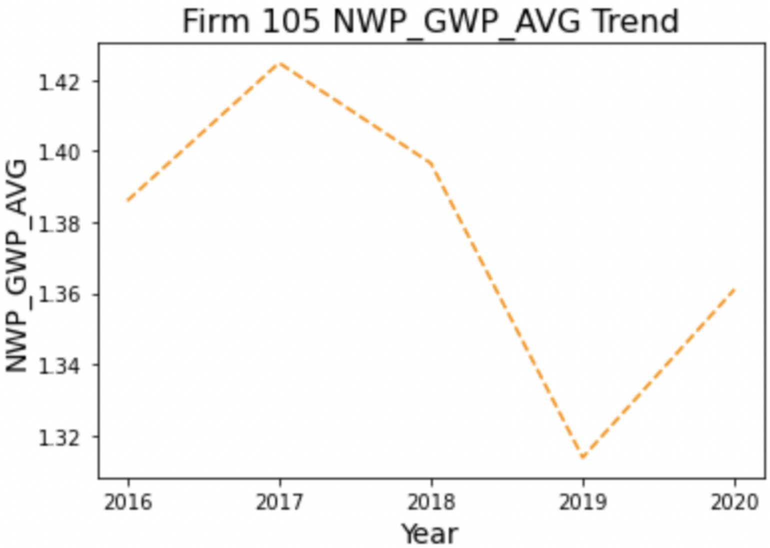
In [18]:
```
# Selecting top 3 firms having NWP/GWP ratio less than 1 i.e. firms which are not insured properly/completely
# Priority order is the GWP i.e.company with more turnover should be given priority
short=df5[df5['NWP_GWP_AVG']<1]
short.sort_values(by=['GWP_AVG'], ascending=False).head(3)
```

Out[18]:

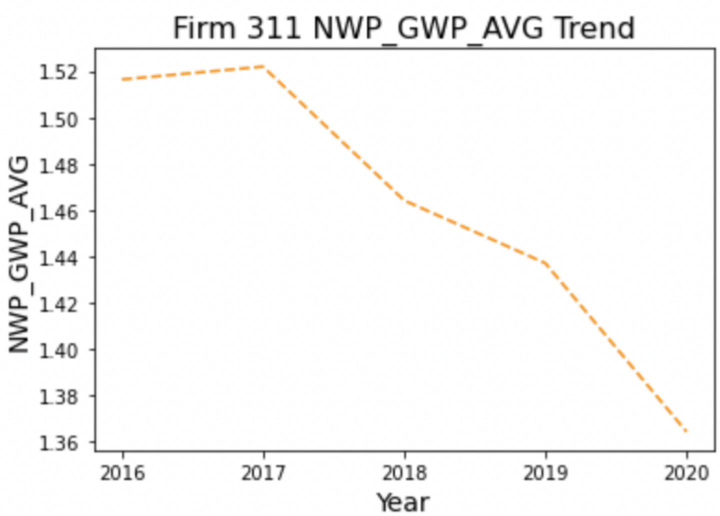
	Firm_Name	GWP_AVG	NWP_GWP_AVG	SCR_ratio_AVG	EOAOL_AVG	NWP_GWP_AVG_STD	SCR_ratio_AVG_STD
3	Firm 4	33857.155951	0.725393	1.476896	19041.638489	2.537884	1.535742
104	Firm 105	14899.454840	0.730077	1.376390	20699.956203	2.537884	1.535742
310	Firm 311	13685.383398	0.393685	1.460739	11662.337006	2.537884	1.535742



2. The NWP/GWP average of Firm 4 is constantly declining since 2019, this implies that the firm doesn't have the complete insurance of the insurers. Since this firm has a GWP average of 33857.155951(Highest), it becomes the most important firm.



3. For Firm 105, there is a variance in their NWP/GWP average but the performance/ ratio dropped significantly low in 2019. Since this firm has a GWP average of 14899.454840 (Second highest), it becomes the second most important firm.



4. For Firm 311, there is a sharp decline in the NWP/GWP average since 2017. This indicates that the insurance of the insurers is constantly decreasing year by year. Since this firm has a GWP average of 13685.383398, it becomes the least important firm out of 3 but significantly important when compared to the total of 325 firms.

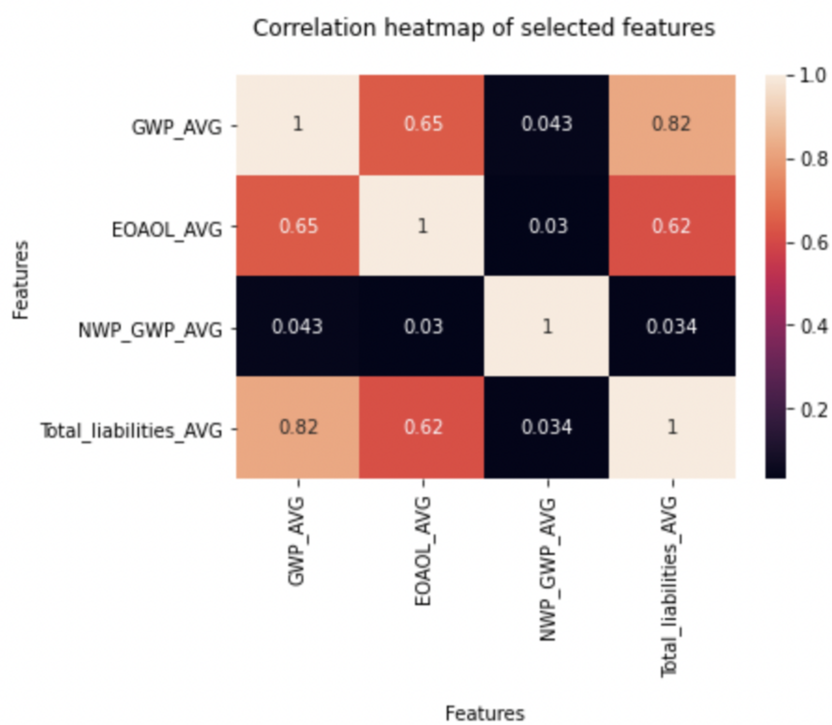
TASK II

Please consider using relevant ML techniques to draw out further insights and present them as an annex to your report.

- **Approach towards the solution:** Building 2 backward step multiple regression machine learning model to predict the value of SCR ratio for future using the current and past values of the metrics.
- **Feature Engineering:**
    - **Correlation Analysis**
      Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables

	GWP_AVG	EOAOL_AVG	NWP_GWP_AVG	Total_liabilities_AVG
GWP_AVG	1.000000	0.646639	0.042675	0.822653
EOAOL_AVG	0.646639	1.000000	0.030023	0.619121
NWP_GWP_AVG	0.042675	0.030023	1.000000	0.034473
Total_liabilities_AVG	0.822653	0.619121	0.034473	1.000000

Here, Total_liabilites_AVG was eliminated/ removed as it is highly correlated to GWP_AVG. Further, the dark colour shows the high correlation between the features and the light colours show less correlation between the features in the below correlation heatmap.



- **Algorithm Selection:**
  Backward Stepwise Multiple Regression: It is a stepwise regression approach that begins with a full (saturated) model and at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data. Also known as Backward Elimination regression.

Feature Engineering

```
In [37]: #Features taken into consideration
         y = (df2['SCR_AVG']).astype(float)
         x1 = (df2['GWP_AVG']).astype(float)
         x2 = (df2['EOAOL_AVG']).astype(float)
         x3 = (df2['NWP_GWP_AVG']).astype(float)
         x4 = (df2['Total_liabilities_AVG']).astype(float)
```

Initially took all the features, x1, x2, x3, and x4 respectively. However, there was a high correlation between x1 and x4, thus removing x4.

Next was running the model with lesser and lesser features in the next iteration. Therefore, in the second iteration it was found that even x3 did not have any impact on the performance. Thus, removed x3 as well.

- **Machine Learning Models:**
  Aim: To predict the value of SCR ratio using EOF_SCR_Avg and SCR_AVG because SCR ratio can be calculated as:

SCR Ratio = EOF / SCR

Features Considered: GWP_AVG, EOAOL_AVG, NWP_GWP_AVG, Total_liabilities_AVG
Features Used: GWP_AVG(x1) , EOAOL_AVG(x2)
Performance Metric
The performance metrics used here is R-squared error, which is a goodness-of-fit measure for linear regression models.This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

Model I : Using EOF_SCR_AVG as a target variable(y)
Performance Metric: The R-squared error value of our model I is 94.06%

```
In [31]: print(f"coefficient of determination: {model.rsquared}")
         print(f"adjusted coefficient of determination: {model.rsquared_adj}")
         print(f"regression coefficients: {model.params}")

coefficient of determination: 0.9406513236135218
adjusted coefficient of determination: 0.9398094984165505
regression coefficients: Intercept   -32.007144
x1            0.059826
x2            0.919898
dtype: float64
```

Here the coefficient values of x1 (0.059826) and x2 (0.919898) depicts that x2 is the most important variable here as 1 unit change in x2 would lead to 91 units change in y.

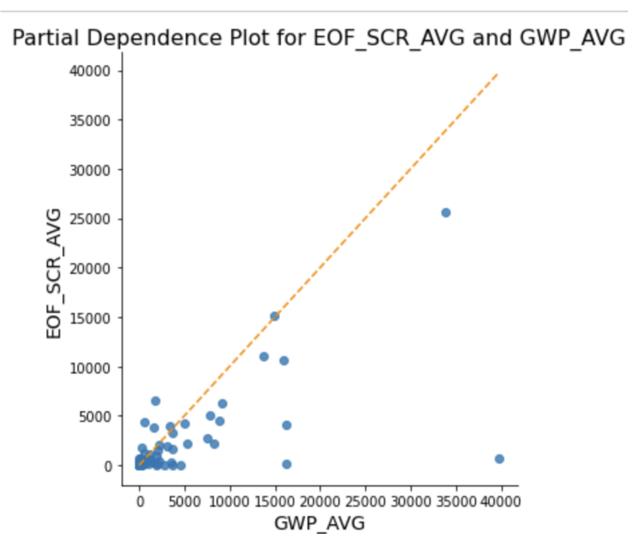
Model I Validation Using F Statistics

Model I Summary and Validation

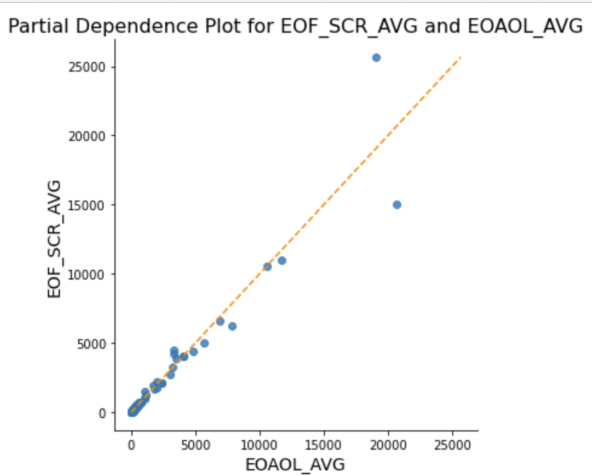
```
[30]: print(model.params)
      print(model.summary())
```

```
Intercept   -32.007144
x1            0.059826
x2            0.919898
dtype: float64
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.941
Model:                            OLS   Adj. R-squared:                  0.940
Method:                 Least Squares   F-statistic:                     1117.
Date:                Fri, 19 Aug 2022   Prob (F-statistic):           3.35e-87
Time:                        01:42:58   Log-Likelihood:                -1150.6
No. Observations:                 144   AIC:                             2307.
Df Residuals:                     141   BIC:                             2316.
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -32.0071     64.607     -0.495      0.621    -159.731      95.717
x1             0.0598      0.015      3.983      0.000       0.030       0.090
x2             0.9199      0.028     33.356      0.000       0.865       0.974
==============================================================================
Omnibus:                      122.487   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15363.477
Skew:                           1.995   Prob(JB):                         0.00
Kurtosis:                      53.445   Cond. No.                     6.45e+03
==============================================================================
```

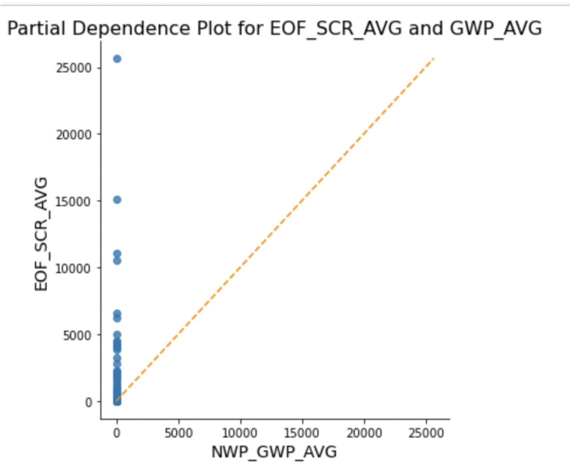
Partial Dependence Plots for Model I



There is a direct relationship between GWP and EOF_SCR. This implies that as the Gross Written Premium/ turnover of the firm increases, their Eligible Own Funds (EOF) also increases .



There is a direct relationship between EOAL and EOF_SCR. This implies that as the Assets increases over Liabilities of the firm, their Eligible Own Funds (EOF) Solvency Capital Requirements (SCR) also increases.



The relationship of reinsuring the insurers (NWP/GWP) can not be identified using EOF_SCR

Model II : Using SCR_AVG as a target variable(y)
Performance Metric: The R-squared error value of our model I is 94.24%

In [52]:
```
print(f"coefficient of determination: {model.rsquared}")
print(f"adjusted coefficient of determination: {model.rsquared_adj}")
print(f"regression coefficients: {model.params}")
```

```
coefficient of determination: 0.9424930343100407
adjusted coefficient of determination: 0.9416773326690484
regression coefficients: Intercept   -50.004103
x1             0.036305
x2             0.626452
dtype: float64
```

Here the coefficient values of x1 (0.036305) and x2 (0.626452) depict that x2 is the most important variable here as 1 unit change in x2 would lead to 62 units change in y.

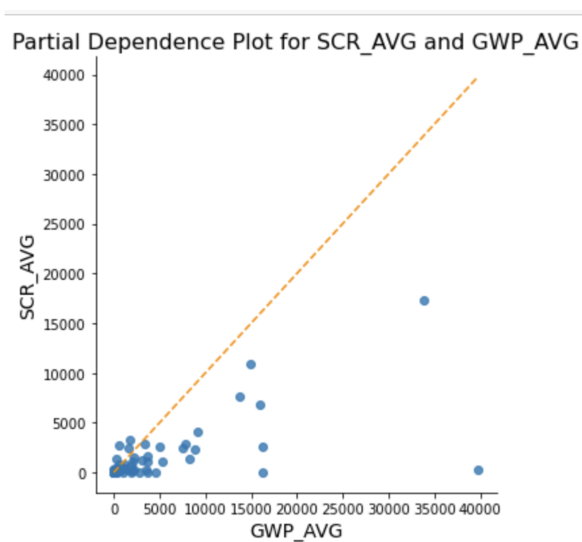
Model II Validation Using F Statistics

Model II Summary and Validation

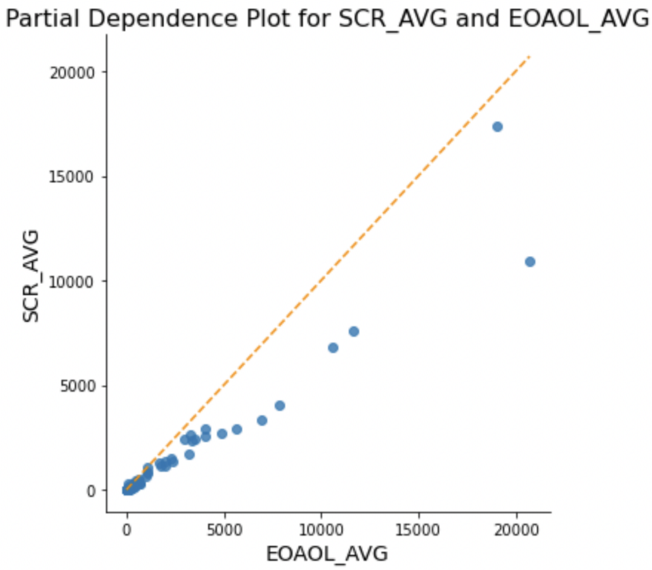
```
In [45]: print(model.params)
         print(model.summary())

Intercept   -50.004103
x1            0.036305
x2            0.626452
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.942
Method:                 Least Squares   F-statistic:                     1155.
Date:                Fri, 19 Aug 2022   Prob (F-statistic):           3.63e-88
Time:                        01:42:59   Log-Likelihood:                -1091.6
No. Observations:                 144   AIC:                             2189.
Df Residuals:                     141   BIC:                             2198.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -50.0041     42.900     -1.166      0.246    -134.814      34.806
x1             0.0363      0.010      3.640      0.000       0.017       0.056
x2             0.6265      0.018     34.210      0.000       0.590       0.663
==============================================================================
Omnibus:                      161.172   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14771.885
Skew:                           3.490   Prob(JB):                         0.00
Kurtosis:                      52.125   Cond. No.                     6.45e+03
==============================================================================
```

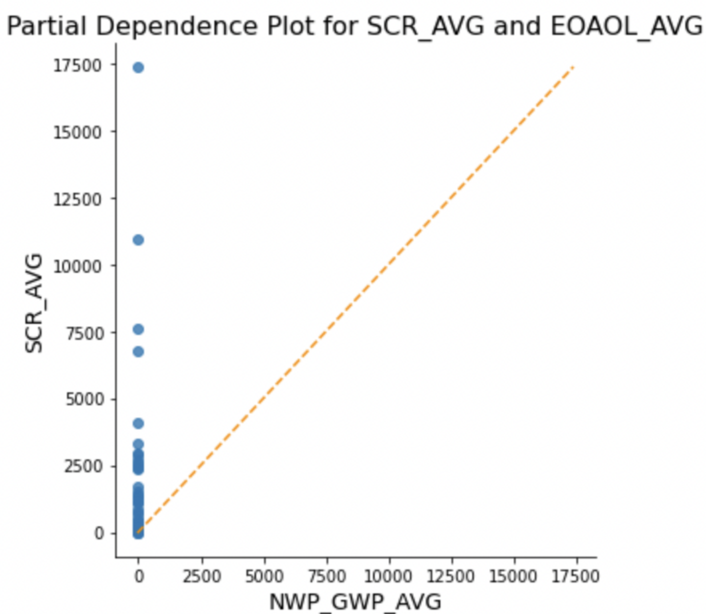
Partial Dependence Plots for Model II



There is a direct relationship between GWP and SCR. This implies that as the Gross Written Premium/ turnover of the firm increases, their Solvency Capital Requirements (SCR) also increases.



There is a direct relationship between EOAL and EOF_SCR. This implies that as the Assets increases over Liabilities of the firm, their Eligible Own Funds (EOF) Solvency Capital Requirements (SCR) also increases.



The relationship of reinsuring the insurers (NWP/ GWP) can not be identified using SCR

- **Further Results/ Insights:**
  1. Trend analysis of Solvency Capital Requirements (SCR) ratio
  2. Formula: SCR Ratio = EOF_SCR / SCR
  3. Using 2 Stepwise Backward Multiple Regression Machine Learning models analysing Eligible Own Fund Solvency Capital Requirements Average (EOF_SCR_AVG) and Solvency Capital Requirements Average (SCR_AVG)

  Using above ML models we can conclude the below points:

  1. There is a direct relationship between GWP and EOF_SCR. This implies that as the Gross Written Premium/ turnover of the firm increases, their Eligible Own Funds (EOF) also increases .
  2. Similarly, there is a direct relationship between GWP and SCR. This implies that as the Gross Written Premium/ turnover of the firm increases, their Solvency Capital Requirements (SCR) also increases.
  3. There is a direct relationship between EOAL and EOF_SCR. This implies that as the Assets increases over Liabilities of the firm, their Eligible Own Funds (EOF) Solvency Capital Requirements (SCR) also increases.
  4. There is a direct relationship between EOAL and SCR. This implies that as the Assets increases over Liabilities of the firm, their Solvency Capital Requirements (SCR) also increases.
  5. The relationship of reinsuring the insurers (NWP/ GWP) can not be identified using SCR or EOF_SCR.

- **Github Repository: https://github.com/sumakshi/Bank-of-England.git**