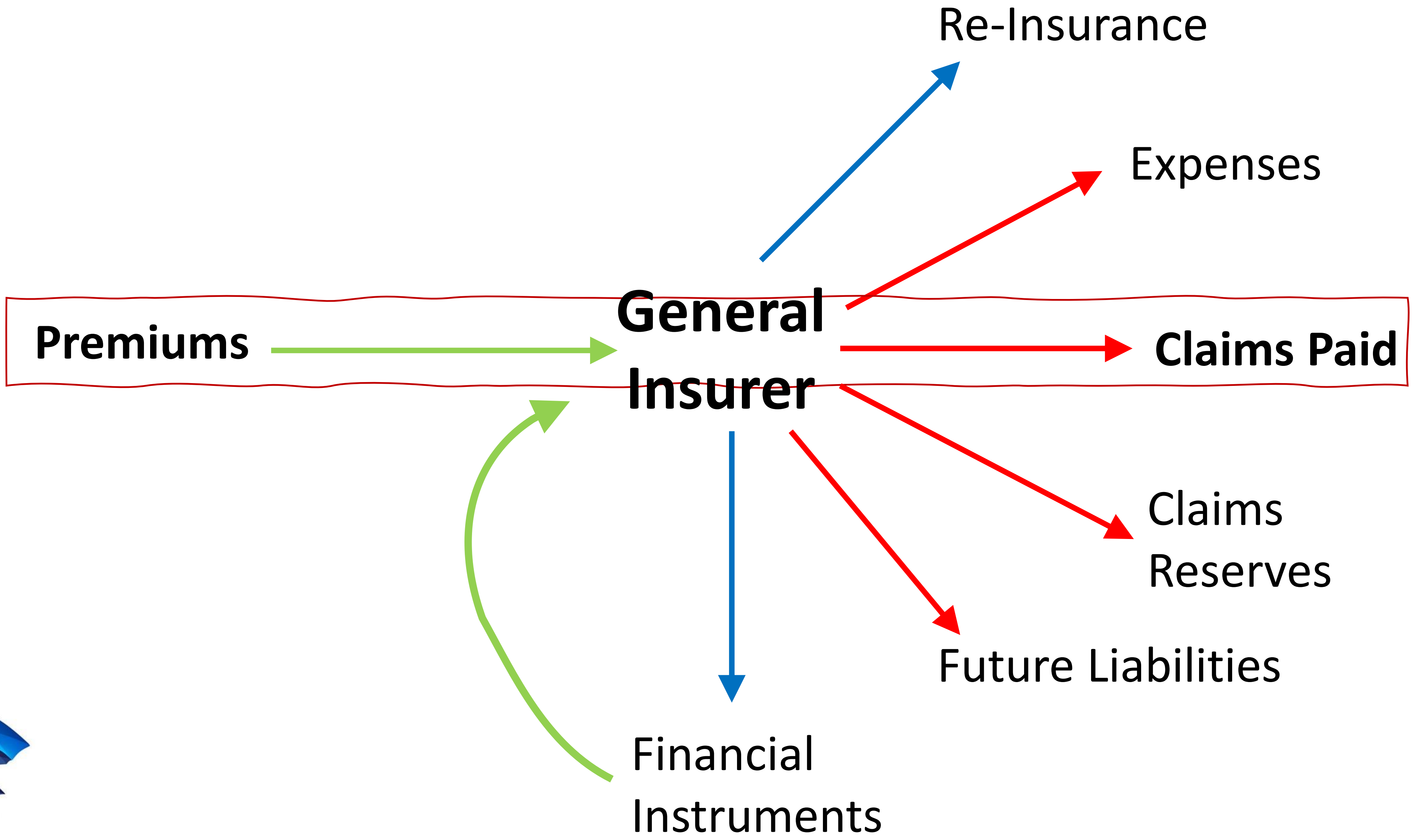
GA DSI-10 CAPSTONE PROJECT

Predicting Whether Insurance Underwriting Gain Will Be Negative

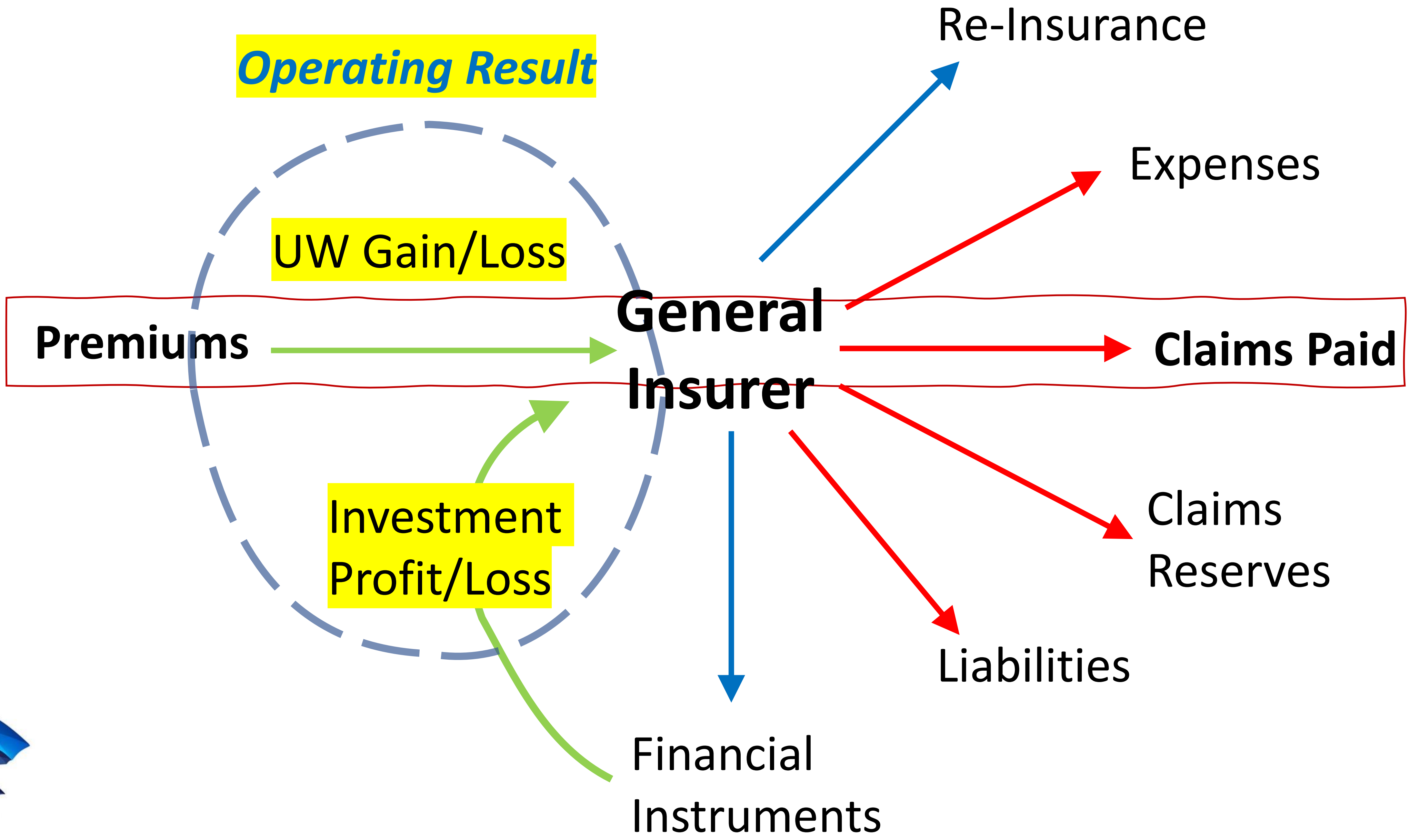
Irwin Wei

5 December 2019

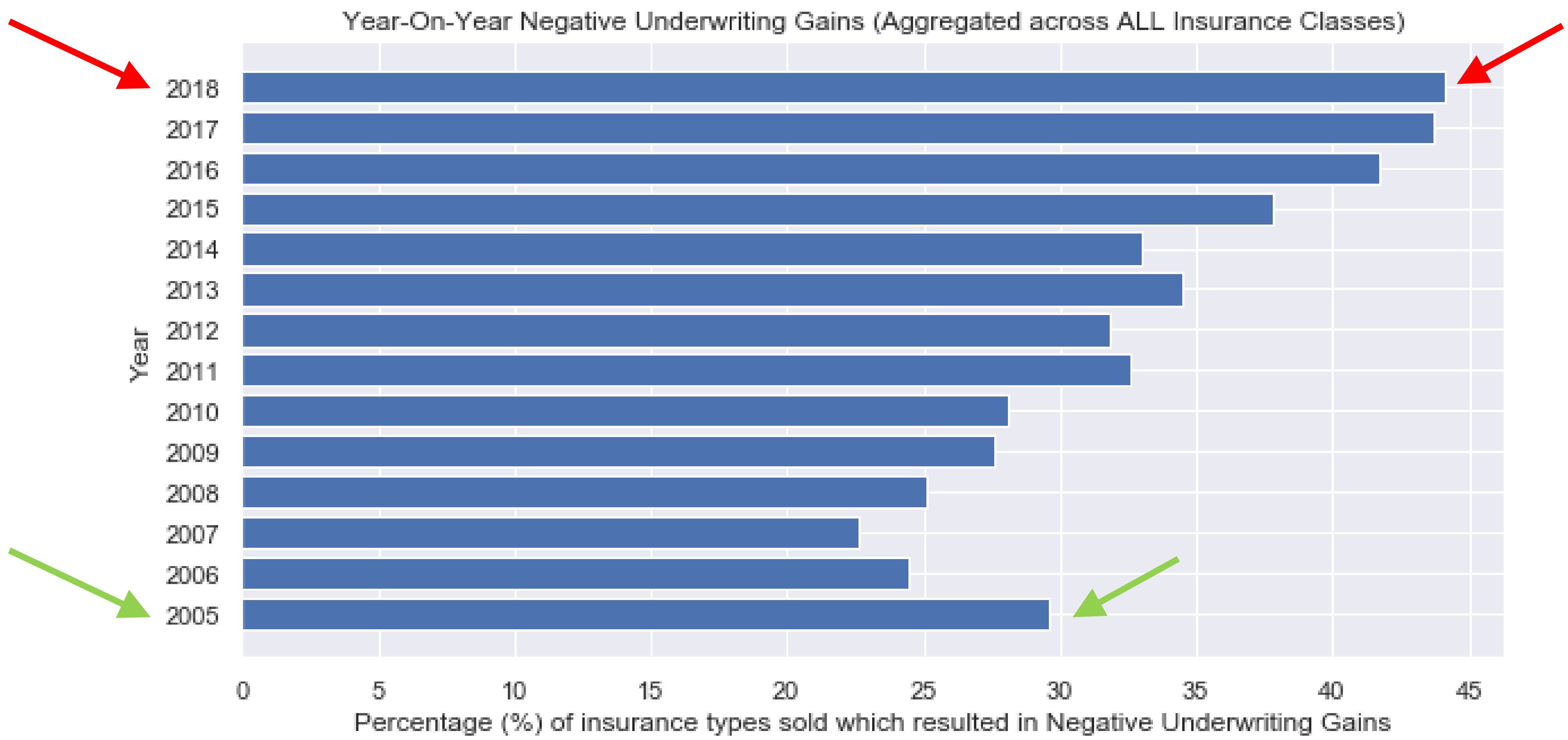
INSURANCE MONEY: INs & OUTs



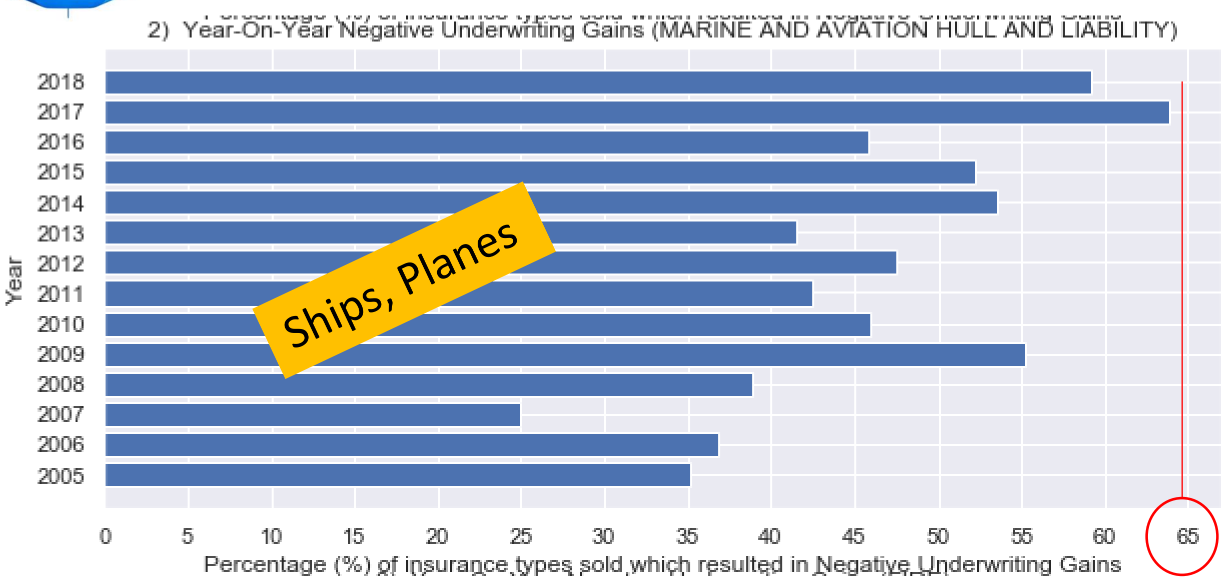
INSURANCE MONEY: INs & OUTs

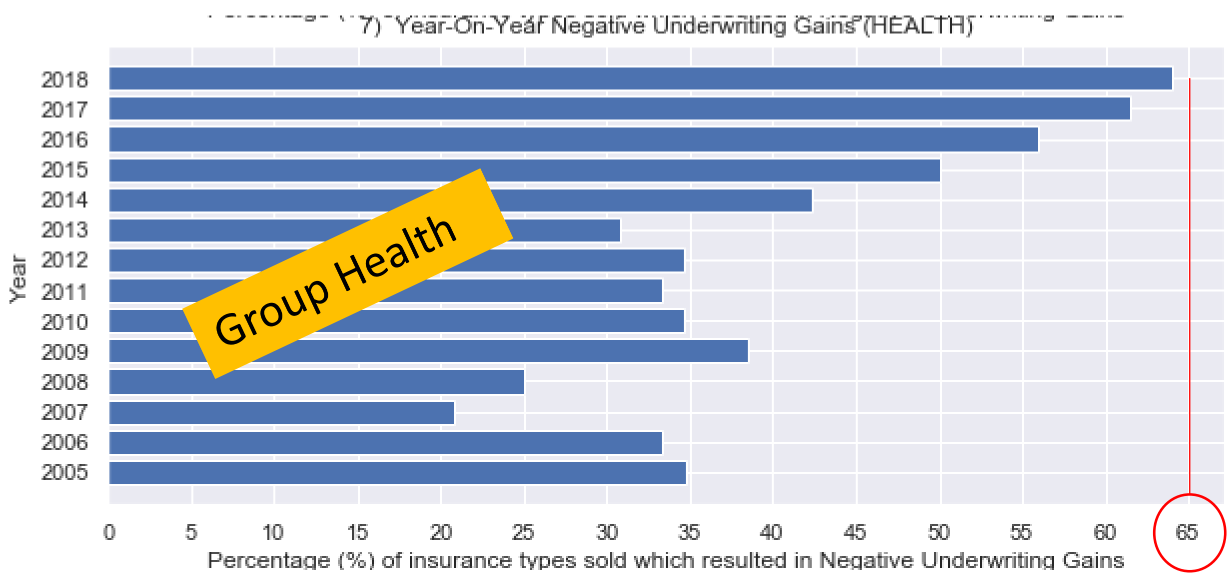


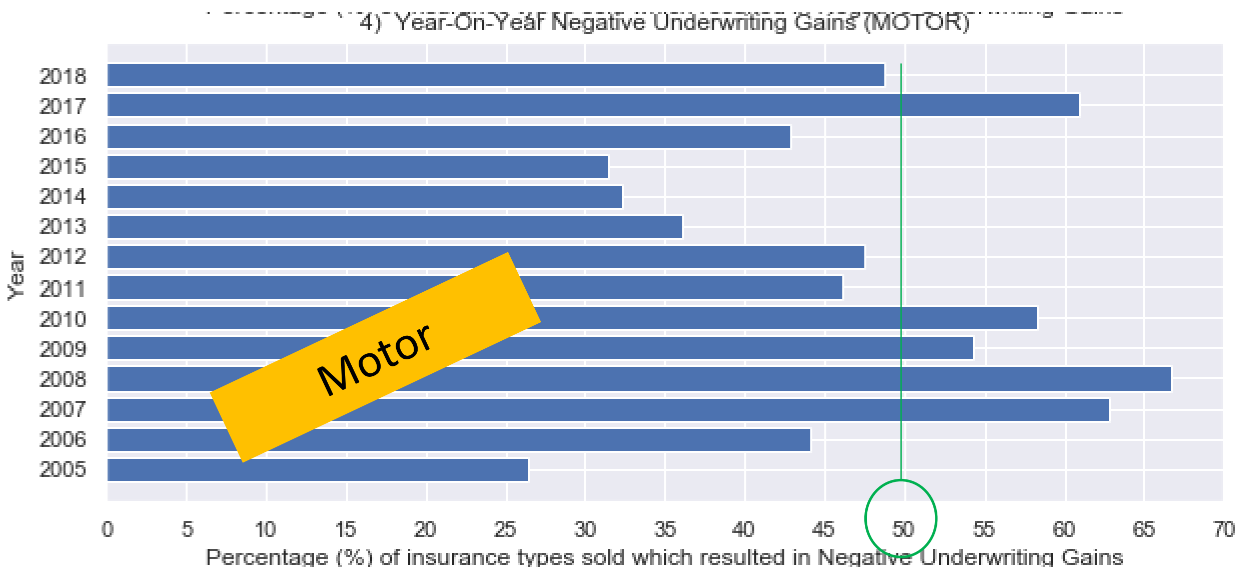
The Rising Trend of Underwriting Losses

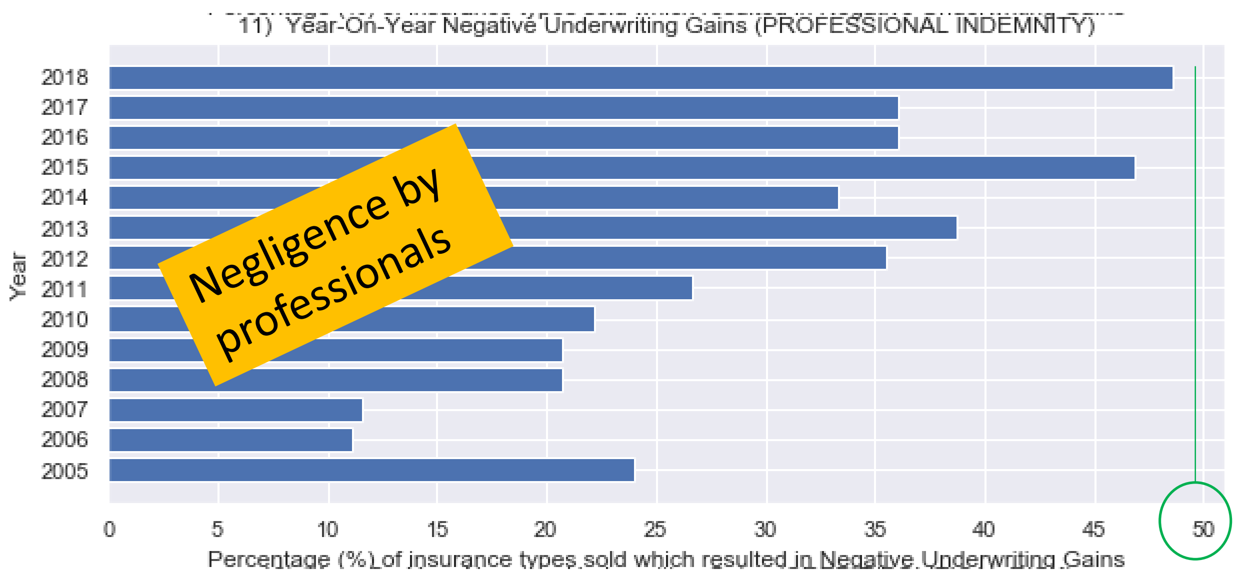


Losing streak …

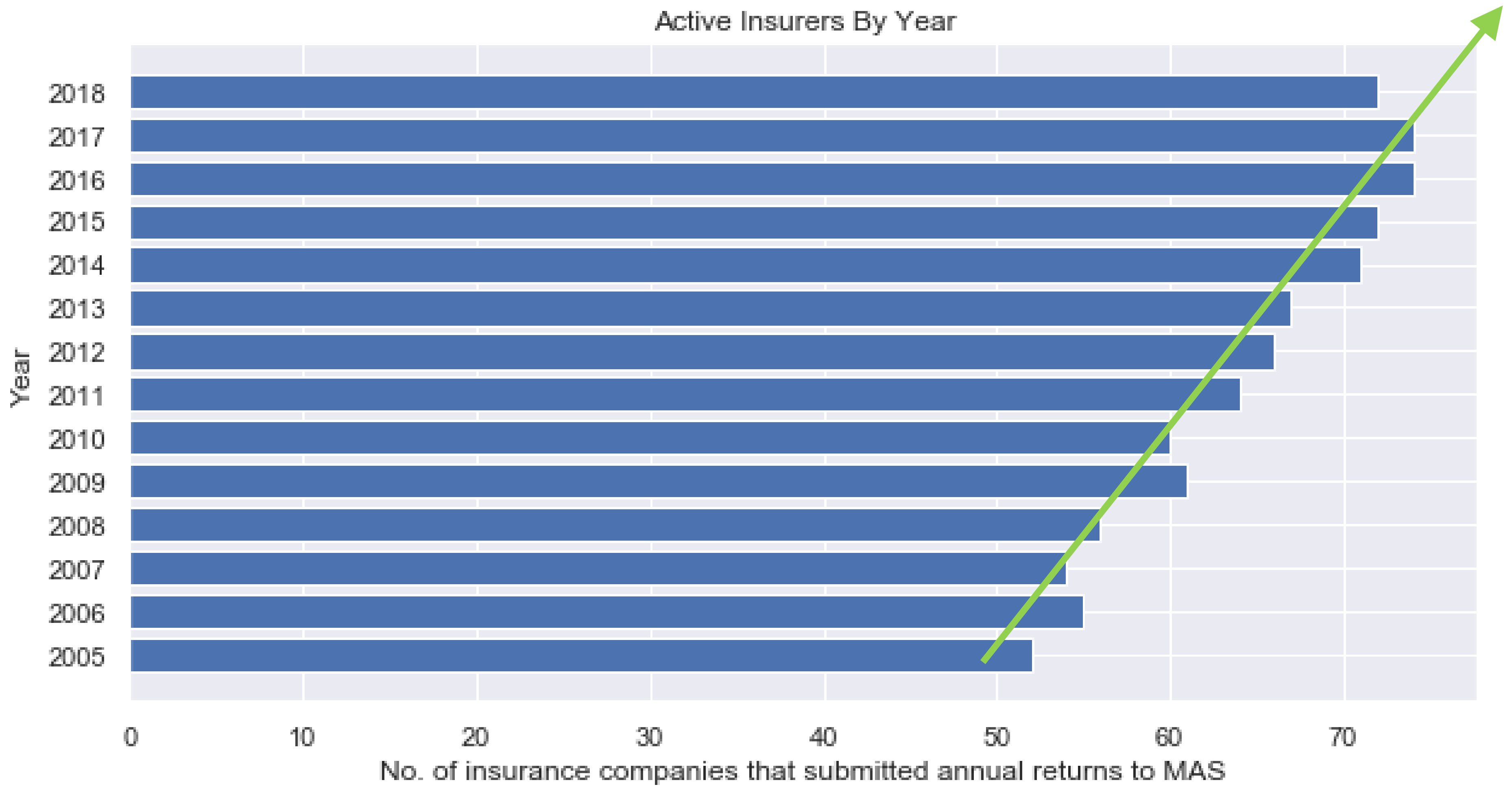








… and yet… More Insurers!



Problem Statement

To implement a proof-of-concept, employing only publicly available data, to predict whether the underwriting performance of any given insurance class is likely to result in an underwriting loss at the end of the current 12-month reporting period.

Value Proposition

With such predictions, underwriters can place more focus on certain insurance classes and review their underwriting approach, and/or take necessary risk management measures such as re-insuring more.

Caveat

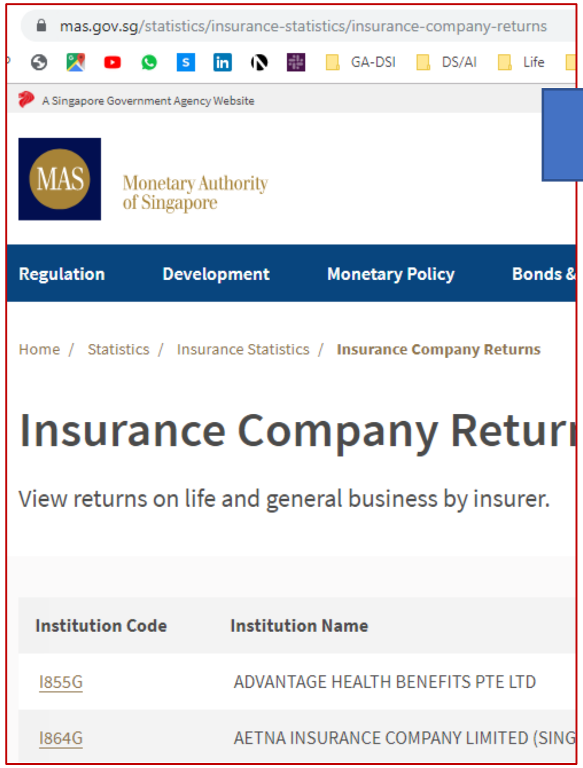
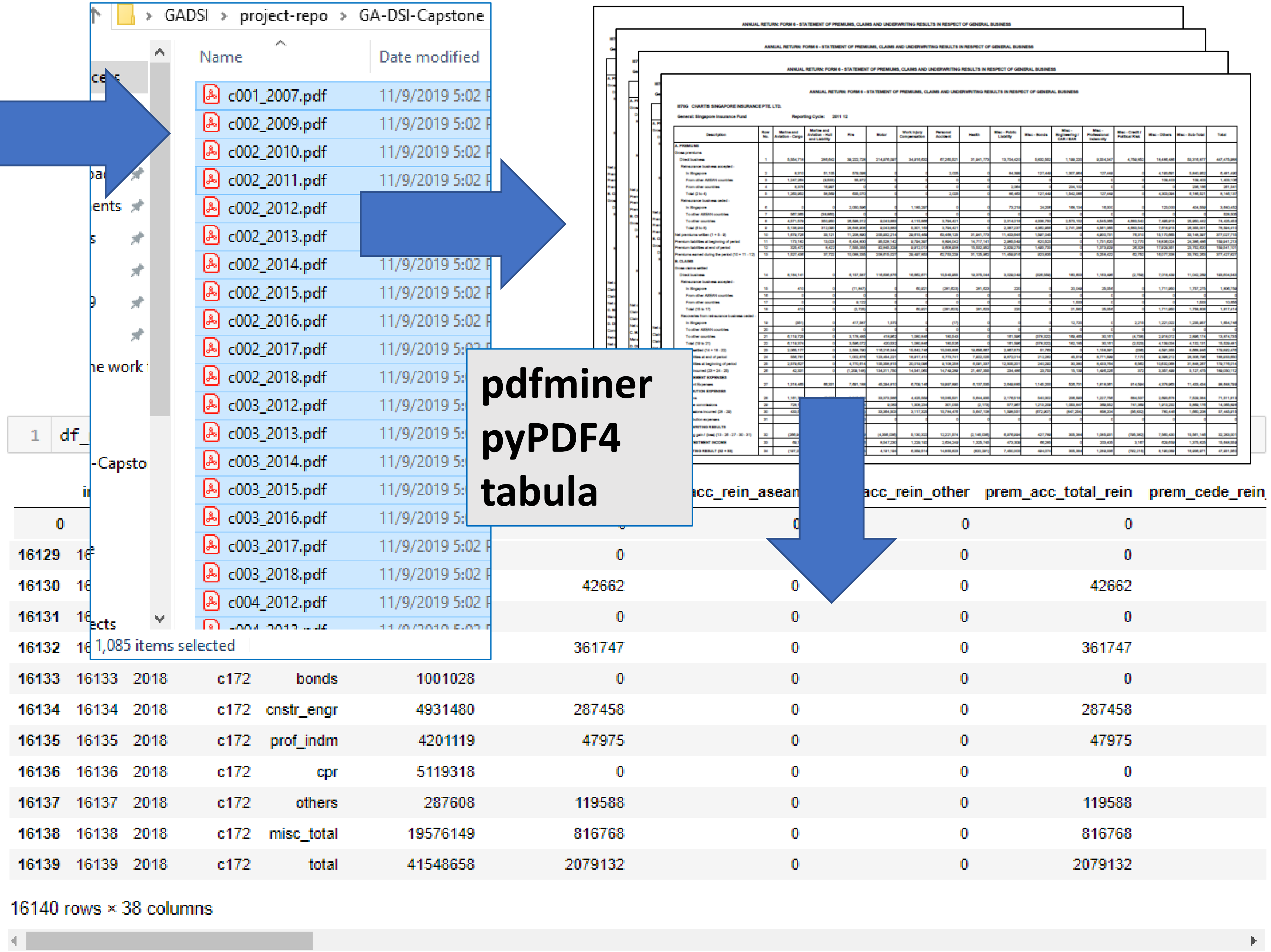
Publicly available data is very coarse. Only shows the start/end state of a 12-month period.

Insurers have much higher quality, high-resolution data concerning underwriting and claims details.

Description	Row No.	Marine and Aviation - Cargo	Marine and Aviation - Hull and Liability	Fire	Motor	Work Injury Compensation	Personal Accident	Health	Misc - Public Liability	Misc - Bonds	Misc - Engineering / CAR / EAR	Misc - Professional Indemnity	Misc - Credit / Political Risk	Misc - Others	Misc - Sub-Total	Total
A. PREMIUMS																
Gross premiums																
Direct business	1	1,459,526	2,947,014	36,747,999	192,134,197	53,595,401	73,154,783	37,503,540	14,951,032	5,196,553	1,622,740	12,455,454	9,333,204	19,854,990	63,313,973	460,456,413
Reinsurance business accepted -																
In Singapore	2	0	349,572	1,123,972	0	0	0	0	632,685	144,034	1,521,022	158,102	0	2,996,321	5,452,164	6,925,708
From other ASEAN countries	3	0	0	9,904	0	0	0	0	0	0	0	0	0	0	0	9,904
From other countries	4	307	407,220	579,129	40,365	11,260	23,248	0	3,274	32,208	660	2,650	1,961	4,758	45,511	1,107,040
Total (2 to 4)	5	307	756,792	1,713,005	40,365	11,260	23,248	0	635,959	176,242	1,521,682	160,752	1,961	3,001,079	5,497,675	8,042,653
Reinsurance business ceded -																
In Singapore	6	0	0	367,875	311	500,024	251,265	(43)	0	0	0	0	0	26,867	26,867	1,146,099
To other ASEAN countries	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
To other countries	8	1,087,998	3,314,779	26,888,179	17,176,901	6,372,411	6,052,257	1,976,062	2,839,529	4,290,824	3,144,640	3,957,346	9,087,467	11,824,476	34,747,286	97,615,873
Total (6 to 8)	9	1,087,998	3,314,779	27,256,054	17,177,212	6,872,435	6,303,523	1,976,019	2,839,529	4,290,824	3,144,640	3,957,346	9,087,467	11,851,143	34,773,953	98,761,972
Net premiums written (1 + 5 - 9)	10	371,835	69,027	11,204,950	174,997,350	46,734,226	66,874,408	35,527,521	12,947,462	1,078,971	(220)	8,658,859	247,698	11,004,926	33,937,695	369,737,093
Premium liabilities at beginning of period	11	48,283	42,860	8,376,684	65,808,041	15,421,829	15,239,680	11,896,811	2,522,735	773,508	0	4,755,206	13,200	15,090,265	23,153,918	159,990,106
Premium liabilities at end of period	12	207,888	53,593	8,271,841	73,993,301	14,320,105	14,674,482	10,159,230	3,260,512	936,000	0	4,952,170	86,040	13,768,323	23,003,048	144,683,488
Premiums earned during the period (10 + 11 - 12)	13	212,230	78,294	11,309,793	166,812,090	47,835,950	67,439,687	37,267,102	12,209,685	916,479	(220)	8,461,896	173,857	12,326,868	34,088,585	385,043,711
B. CLAIMS																
Gross claims settled																
Direct business	14	24,006	595,197	10,807,119	109,041,158	23,334,264	16,820,483	19,036,153	1,290,907	977,360	36,648	1,971,299	453,098	3,826,314	8,325,657	197,764,058
Reinsurance business accepted -																
In Singapore	15	0	0	0	0	0	0	0	0	11,783	421,783	9,690	0	0	443,256	443,256
From other ASEAN countries	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
From other countries	17	0	0	5,644	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]	0	5,644
Total (15 to 17)	18	0	0	5,644	0	0	0	0	0	11,783	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	446,900
Recoveries from reinsurance business ceded -																
In Singapore	19	0	0	82,531	625	0	0	0	0	(2,196)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	81,301
To other ASEAN countries	20	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0
To other countries	21	18,006	592,992	10,112,656	57,855	1,454,806	465,162	107,177	317,959	969,281	458,401	368,422	447,078	12,209	2,593,372	15,401,628
Total (19 to 21)	22	18,006	592,992	10,195,187	58,480	1,454,806	465,162	107,177	317,959	967,083	458,401	[illegible]	447,078	[illegible]	[illegible]	15,482,959
Net claims settled (14 + 18 - 22)	23	6,002	2,205	617,576	108,982,678	21,879,676	16,155,321	18,928,975	942,948	22,080	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	172,749,967
Claims liabilities at end of period	24	29,599	103,277	1,487,829	121,496,439	41,762,917	13,818,106	8,373,442	8,892,889	465,685	[illegible]	8,180,944	226,480	3,769,823	21,296,260	208,637,869
Claims liabilities at beginning of period	25	8,417	57,931	1,043,026	119,545,941	34,818,724	14,594,125	6,562,127	6,403,887	606,715	11,375	8,827,785	101,673	5,890,230	21,843,465	196,464,756
Net claims incurred (23 + 24 - 25)	26	27,184	47,551	1,062,179	110,933,174	28,822,869	15,369,302	20,740,290	3,431,950	(100,937)	19,451	825,736	130,846	1,493,325	5,900,361	182,922,910
C. MANAGEMENT EXPENSES																
Management Expenses	27	296,507	890,072	7,811,880	42,572,178	10,867,975	18,585,196	9,509,653	3,180,037	1,323,067	637,496	2,557,753	1,892,570	4,593,191	14,184,124	104,477,367
D. DISTRIBUTION EXPENSES																
Commissions	28	254,422	516,922	2,702,179	29,623,007	5,975,028	17,899,719	6,926,463	2,942,921	693,285	252,573	1,719,314	1,588,145	2,412,588	9,608,826	73,506,586
Reinsurance commissions	29	665,047	900,721	4,921,557	361,282	938,160	979,590	149,138	674,585	1,617,275	1,072,462	252,014	2,901,692	3,182,739	9,700,768	18,616,283
Net commissions incurred (28 - 29)	30	(410,625)	(383,799)	(2,219,378)	29,261,725	5,036,868	16,920,129	6,777,325	2,268,336	(923,990)	(819,889)	1,467,300	(1,313,548)	(770,151)	(91,942)	54,890,303
Other distribution expenses	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E. UNDERWRITING RESULTS																
Underwriting gain / (loss) (13 - 26 - 27 - 30 - 31)	32	299,164	(275,530)	4,655,332	4,045,013	3,108,238	16,585,058	239,634	3,349,382	616,319	162,732	3,511,117	(536,011)	7,010,503	14,116,022	42,753,131
F. NET INVESTMENT INCOME	33	9,492	2,273	266,033	4,467,223	1,193,002	1,707,130	906,924	330,515	37,540	0	221,036	6,323	380,921	866,340	9,438,417

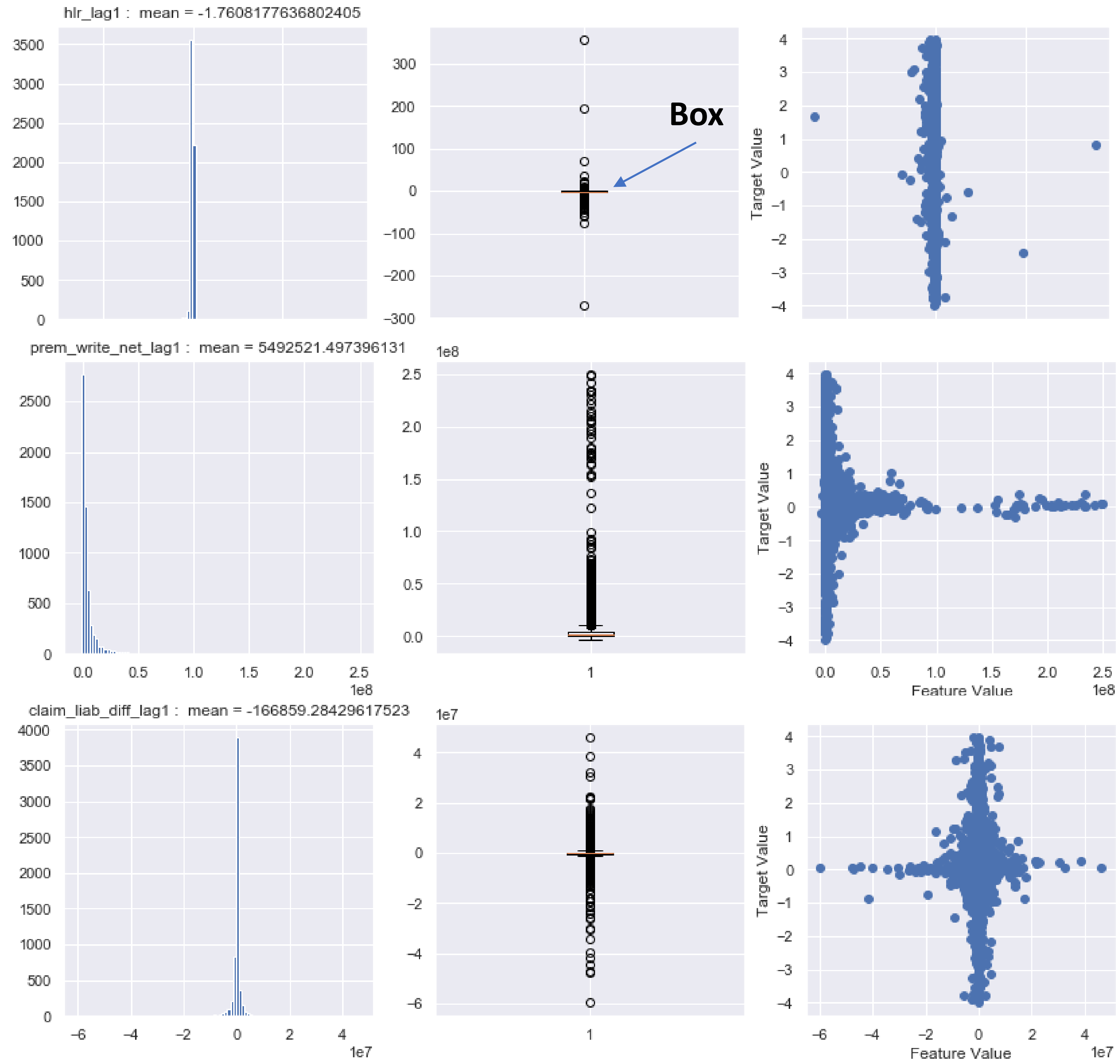
Annual Returns Submitted to MAS – Table “FORM 6 (SIF)”

Get data ...

mas.gov.sg/statistics/insurance-statistics/insurance-company-returns
GA-DSI
DS/AI
Life
A Singapore Government Agency Website
MAS
Monetary Authority of Singapore
Regulation
Development
Monetary Policy
Bonds &
Home / Statistics / Insurance Statistics / Insurance Company Returns
Insurance Company Retur
View returns on life and general business by insurer.
Institution Code
Institution Name
I855G
ADVANTAGE HEALTH BENEFITS PTE LTD
I864G
AETNA INSURANCE COMPANY LIMITED (SING
GADSI › project-repo › GA-DSI-Capstone
Name
Date modified
c001_2007.pdf
c002_2009.pdf
c002_2010.pdf
c002_2011.pdf
c002_2012.pdf
c002_2013.pdf
c002_2014.pdf
c002_2015.pdf
c002_2016.pdf
c002_2017.pdf
c002_2018.pdf
c003_2012.pdf
c003_2013.pdf
c003_2014.pdf
c003_2015.pdf
c003_2016.pdf
c003_2017.pdf
c003_2018.pdf
c004_2012.pdf
11/9/2019 5:02
1,085 items selected
pdfminer
pyPDF4
tabula
acc_rein_asean
acc_rein_other
prem_acc_total_rein
prem_cede_rein
16133 16133 2018 c172 bonds 1001028 0 0 0 0
16134 16134 2018 c172 cnstr_engr 4931480 287458 0 0 287458
16135 16135 2018 c172 prof_indm 4201119 47975 0 0 47975
16136 16136 2018 c172 cpr 5119318 0 0 0 0
16137 16137 2018 c172 others 287608 119588 0 0 119588
16138 16138 2018 c172 misc_total 19576149 816768 0 0 816768
16139 16139 2018 c172 total 41548658 2079132 0 0 2079132
16140 rows × 38 columns

Non-Ideal data



~ 9,000 rows

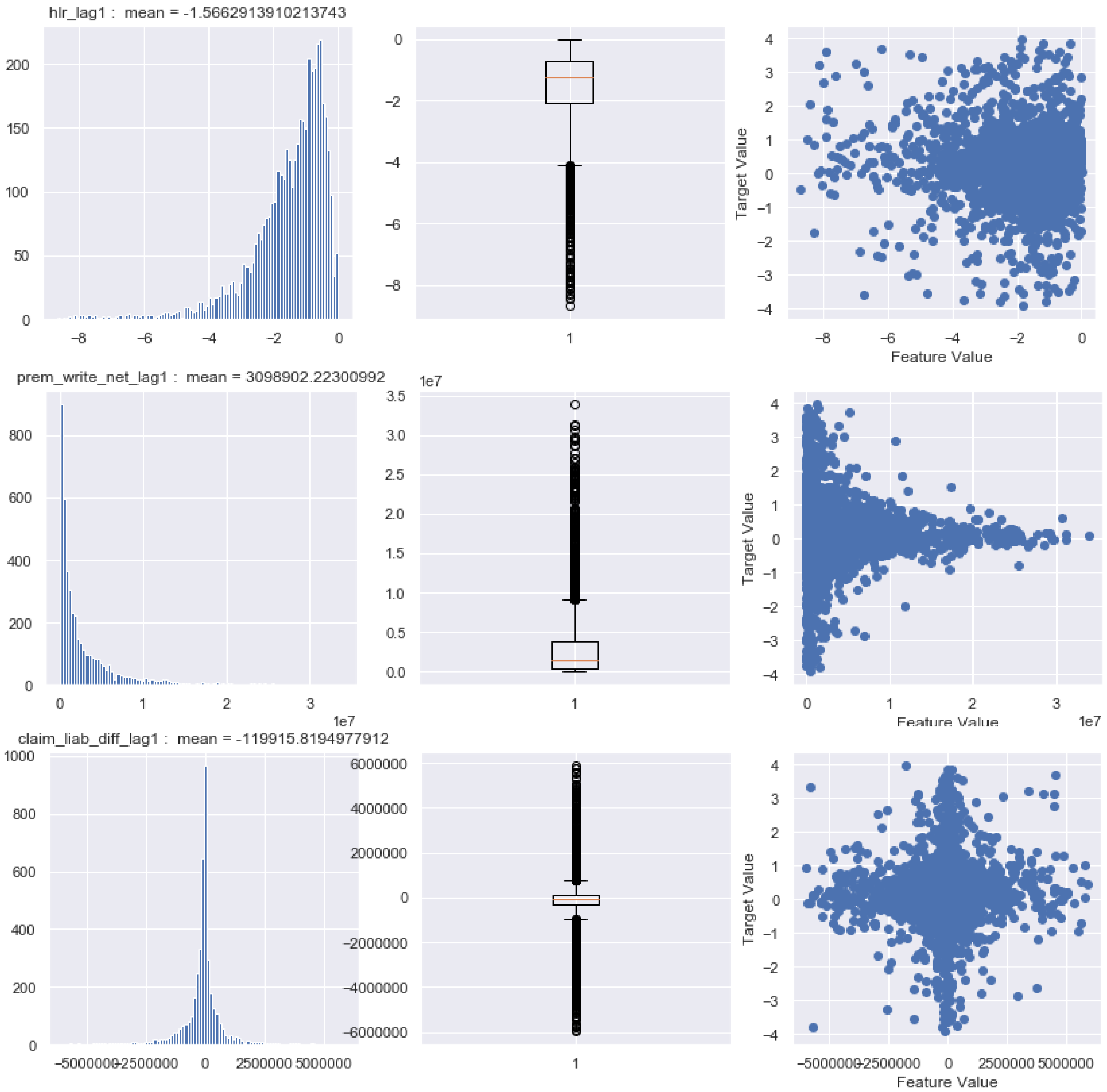
FEATURE ENGINEERING

- Calculated ratios in order to normalize data
- Used exponential weighted means on time-lag data (up to 5 years' history)
- Reduced to 5 features, then built them up again
- +/- polarity of features depending on whether it represented incoming/outgoing money

After trimming away Extreme Outliers

&

Feature Engineering



~ 4,400 rows

REGRESSION MODELS

- Linear Regression
- Decision Tree
- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boosting

Feature Re-Engineering

UNSUPERVISED LEARNING

- K-Means
- DBSCAN
- Hierarchical

No meaningful or interpretable clusters…

REGRESSION MODELS

- Linear Regression
- Decision Tree
- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boosting

Feature Re-Engineering

Best R2 Score => 0.53 (only)

UNSUPERVISED LEARNING

- K-Means
- DBSCAN
- Hierarchical

No meaningful or interpretable clusters…

CLASSIFICATION

Positive Case (1) ➔ Underwriting Gain < 0

Negative Case (0) ➔ Underwriting Gain >= 0

```
In [8]:  1  df['classification'].value_counts(normalize=True).sort_index()
executed in 34ms, finished 16:49:39 2019-12-04

Out[8]: 0    0.680699
        1    0.319301
        Name: classification, dtype: float64
```

Baseline Accuracy is 0.68

CLASSIFICATION MODELS

- Logistic Regression
- Decision Tree
- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boosting
- K Nearest Neighbors
- Support Vector Machine

METRICS:

Recall

Accuracy

CLASSIFICATION MODELS

- Logistic Regression
- Decision Tree
- Random Forest
- Extra Trees
- Ada Boost
- Gradient Boosting
- K Nearest Neighbors
- Support Vector Machine

METRICS:

Recall

Accuracy

6.1 Logistic Regression Classifier

After Tuning

Baseline Accuracy is 0.68

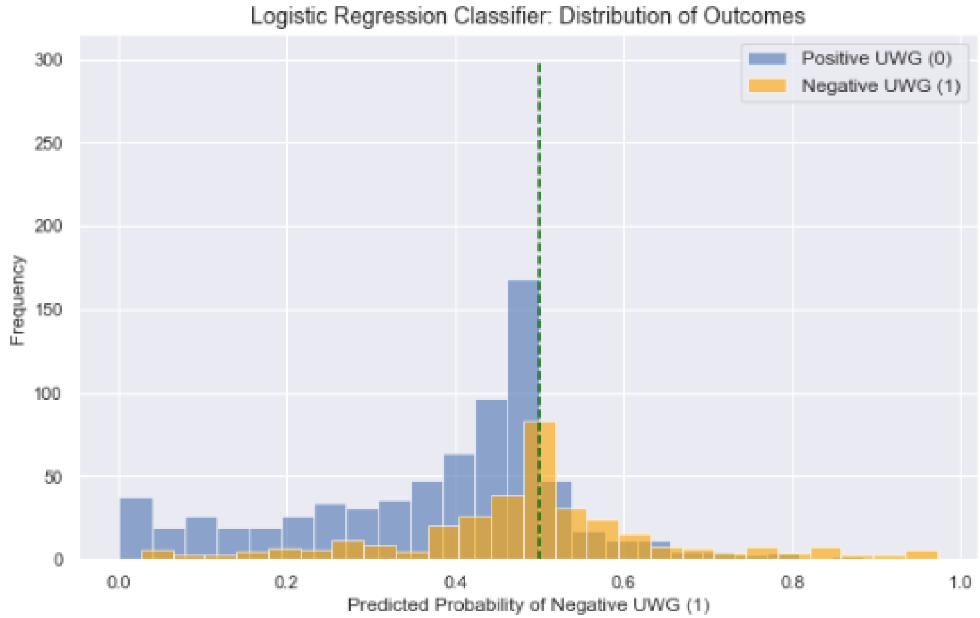
In [13]:

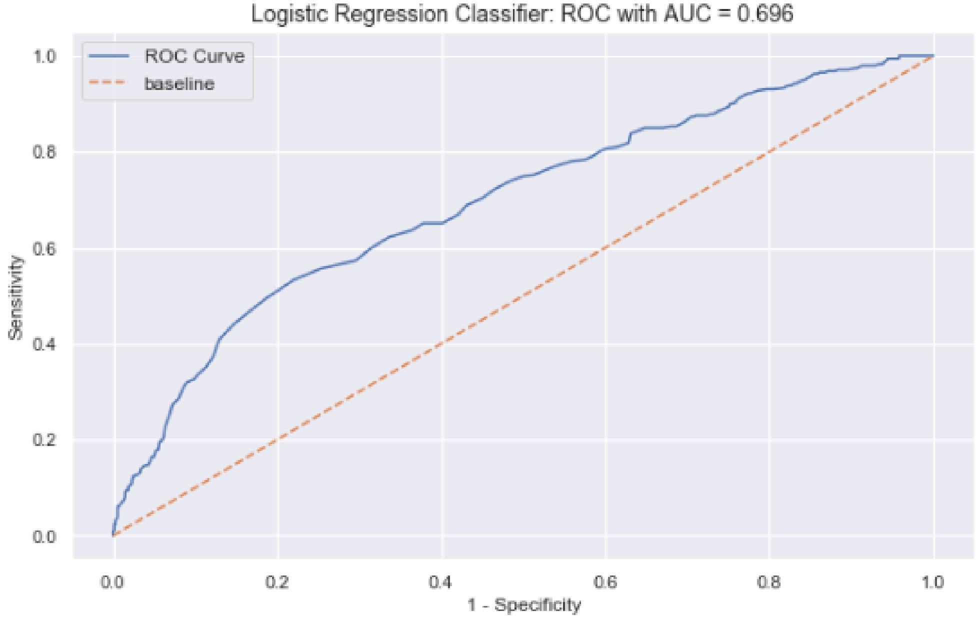
```
# lr = LogisticRegression(fit_intercept=False,C=1.0,tol
lr = LogisticRegression(solver='lbfgs')
lr.fit(X_train,y_train)
print('Score(train/test):',lr.score(X_train, y_train),'
# print('Score (test):\t',lr.score(X_test, y_test))
print('\n=== Classification Report  ===================
# predict & evaluate
predictions = lr.predict(X_test)
print(classification_report(y_test,predictions,target_n
```

executed in 85ms, finished 16:49:39 2019-12-04

```
Score(train/test): 0.7239263803680982 , 0.7230910763569457

=== Classification Report  ================================
                   precision    recall  f1-score   support

Positive UWG (0)       0.76      0.86      0.81       740
Negative UWG (1)       0.59      0.44      0.50       347

        accuracy                           0.72      1087
       macro avg       0.68      0.65      0.65      1087
    weighted avg       0.71      0.72      0.71      1087
```

```
Score(train/test): 0.7220858895705522 , 0.7230910763569457

=== Classification Report  ====================================
                   precision    recall  f1-score   support

Positive UWG (0)       0.77      0.84      0.81       740
Negative UWG (1)       0.58      0.46      0.52       347

        accuracy                           0.72      1087
       macro avg       0.68      0.65      0.66      1087
    weighted avg       0.71      0.72      0.71      1087
```





{'C': 1.0, 'fit_intercept': True, 'max_iter': 5000, 'penalty': 'l2', 'solver': 'sag', 'tol': 0.1}

6.4 Extra Trees Classifier

After Tuning

Baseline Accuracy is 0.68

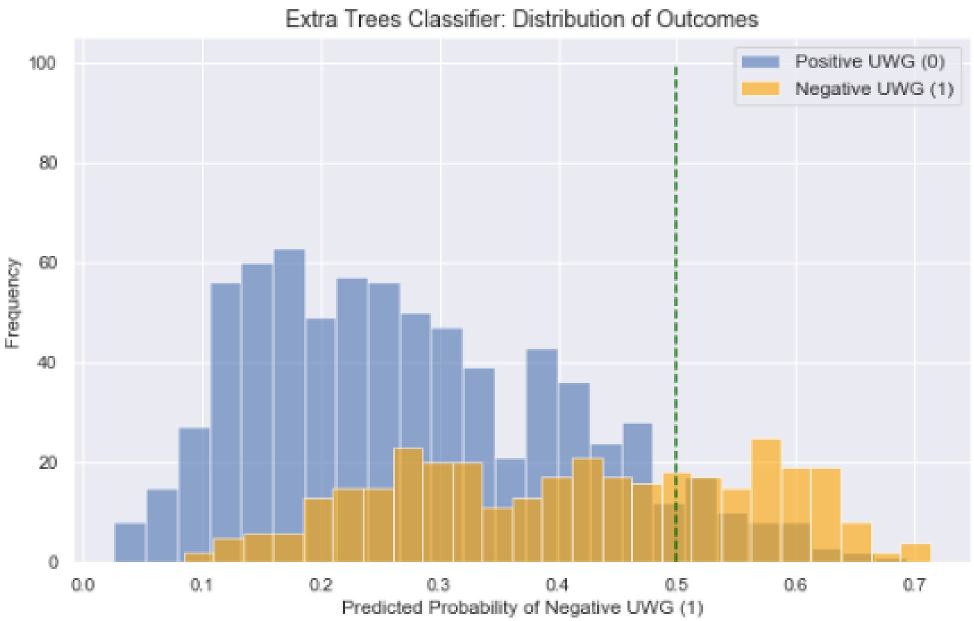
In [23]:

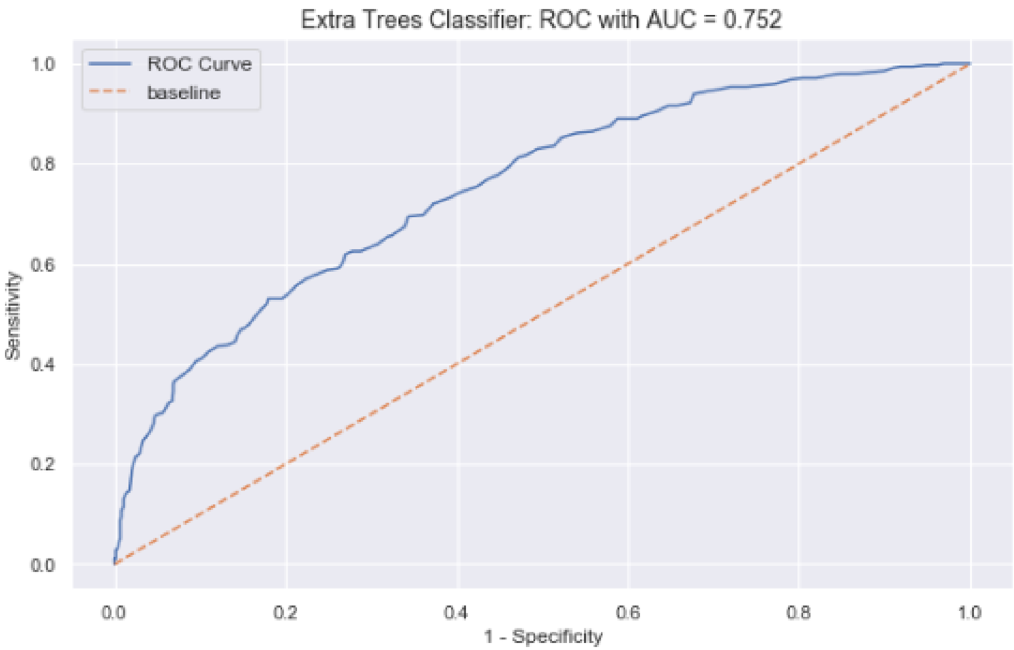
```
etcf = ExtraTreesClassifier(bootstrap=True,oob_score=T
etcf.fit(X_train,y_train)
# Evaluate model.
print('Score(train/test):',etcf.score(X_train, y_trair
# predict & evaluate
predictions = etcf.predict(X_test)
print('\n=== Classification Report  =================
print(classification_report(y_test,predictions,target_
```

executed in 887ms, finished 16:50:16 2019-12-04

```
Score(train/test): 1.0 , 0.7332106715731371

=== Classification Report  ================================
                  precision    recall  f1-score   support

Positive UWG (0)       0.75      0.91      0.82       740
Negative UWG (1)       0.64      0.37      0.47       347

        accuracy                           0.73      1087
       macro avg       0.70      0.64      0.64      1087
    weighted avg       0.72      0.73      0.71      1087
```

```
Score(train/test): 0.807668711656417 , 0.7396504139834407

=== Classification Report  ===============================
                  precision    recall  f1-score   support

Positive UWG (0)       0.75      0.93      0.83       740
Negative UWG (1)       0.70      0.33      0.45       347

        accuracy                           0.74      1087
       macro avg       0.72      0.63      0.64      1087
    weighted avg       0.73      0.74      0.71      1087
```





{'bootstrap': True, 'max_depth': 31, 'min_samples_split': 27, 'n_estimators': 150, 'oob_score': True}

Comparison

Logistic Regressor

```
Score(train/test): 0.7220858895705522 , 0.7230910763569457

=== Classification Report  ==================================
                  precision    recall  f1-score   support

Positive UWG (0)       0.77      0.84      0.81       740
Negative UWG (1)       0.58      0.46      0.52       347

        accuracy                           0.72      1087
       macro avg       0.68      0.65      0.66      1087
    weighted avg       0.71      0.72      0.71      1087
```

Extra Trees

```
Score(train/test): 0.8076687116564417 , 0.7396504139834407

=== Classification Report  ==================================
                  precision    recall  f1-score   support

Positive UWG (0)       0.75      0.93      0.83       740
Negative UWG (1)       0.70      0.33      0.45       347

        accuracy                           0.74      1087
       macro avg       0.72      0.63      0.64      1087
    weighted avg       0.73      0.74      0.71      1087
```

Conclusion

- *Whilst Extra Trees did slightly better in accuracy, Logistic Regression performed much better on Recall ← and that's the most important metric for this problem statement.*

- *Using coarse-grained, low-resolution data from the annual returns alone, the classification model was already able to out-perform the baseline accuracy for classification. Extrapolating further, should insurers'* ***underwriting and claims data*** *be available for analysis, it is certain that far greater insights can be extracted.*