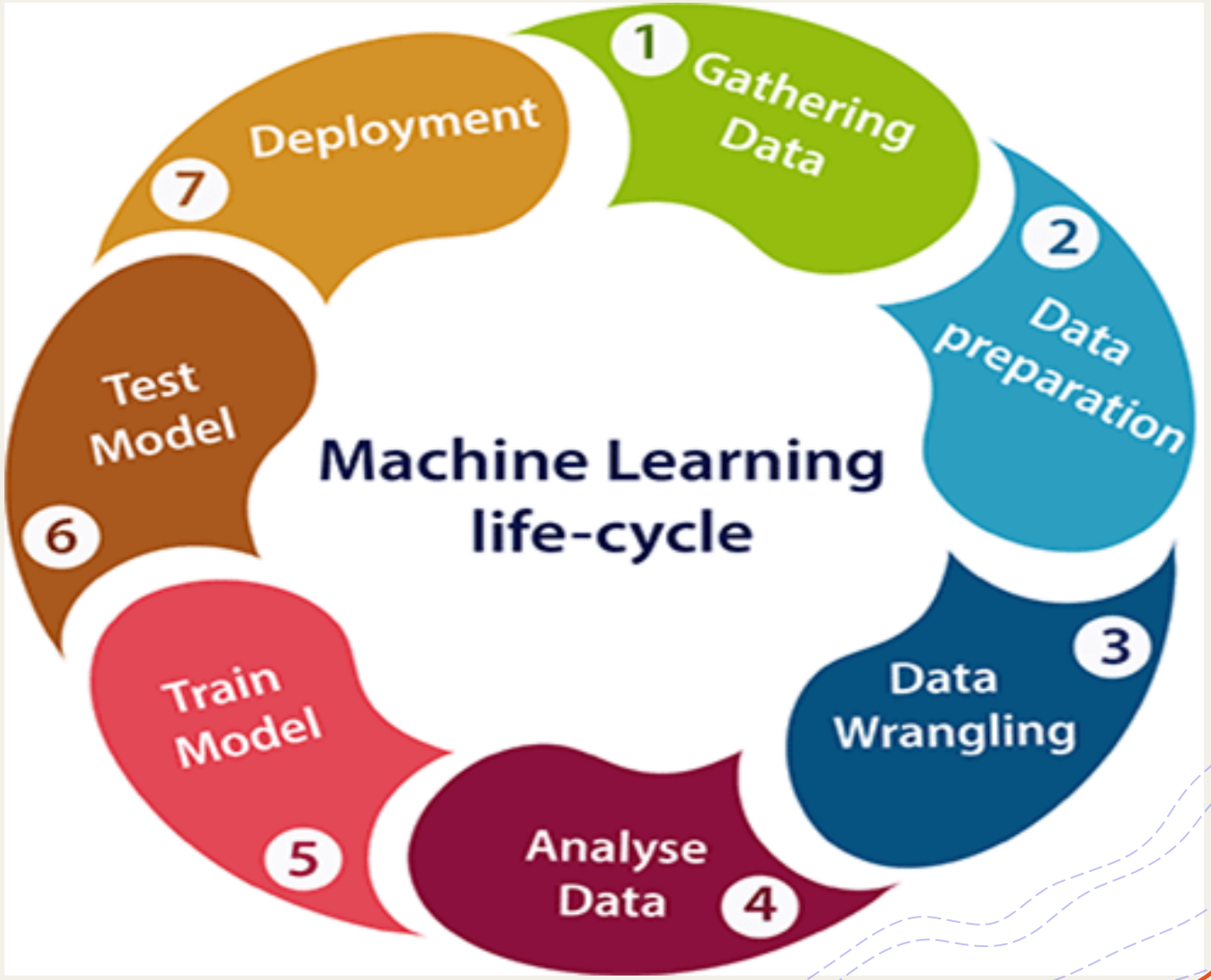
Milestone 4: Feature Selection And Improvements in The Machine Learning Lifecycle

Albina Cako & Joshua Dalphy



https://www.google.com/search?q=machine+learning+lifecycle&sxsrf=ALeKk01bUbJphNoNwKuiHiUzxclCM-RnNQ:1611175855828&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiY-7nksavuAhXpFlkFHSggB34Q_AUoAXoECB4QAw&biw=1033&bih=899#imgrc=fCY4CnUhNfEF_M

Objectives

+ Apply the feature selection methods learned in class to our dataset in order to improve our baseline model's performance.
+ Analyze the effect the number of selected features has on model performance.
+ Compare the results of our multiple iterations in order to identify the best process and model.

Quick Review

1. Large movie review dataset containing 3.5k positive/negative reviews and sentiment (1 = positive & 0 = negative).
2. A detailed exploration of the data was undertaken to better understand the dataset.
3. Feature extraction methods were studied and BOW TFIDF was applied.
4. Baseline models were generated using Naïve Bayes, Logistic Regression and Random Forest.
5. Ensemble learning was conducted using stacking, boosting and bagging methods.
6. All models were evaluated and compared using various metrics and the tune logistic regression model performed best.

Improving Model Performance

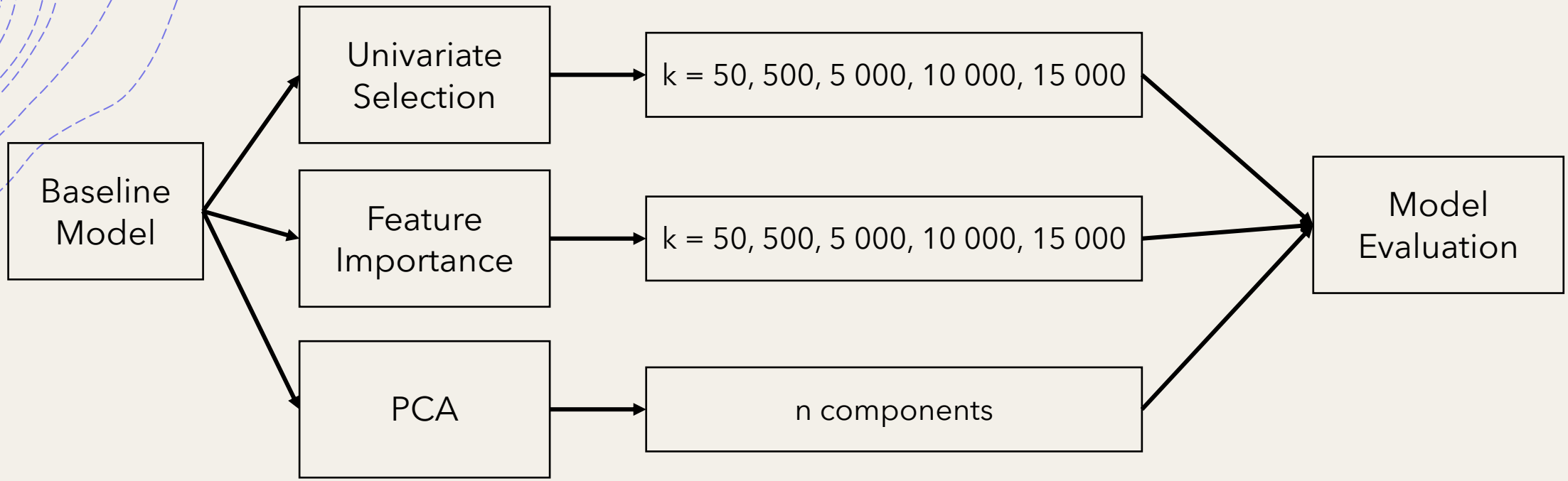
+ Common methods of improving model performance are:

+ Adding more data
+ Feature Engineering
+ Feature Selection

+ Ensemble Learning
+ Cross Validation
+ Tuning

Reference: https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results

Methodology



<u>This process was done twice:</u>

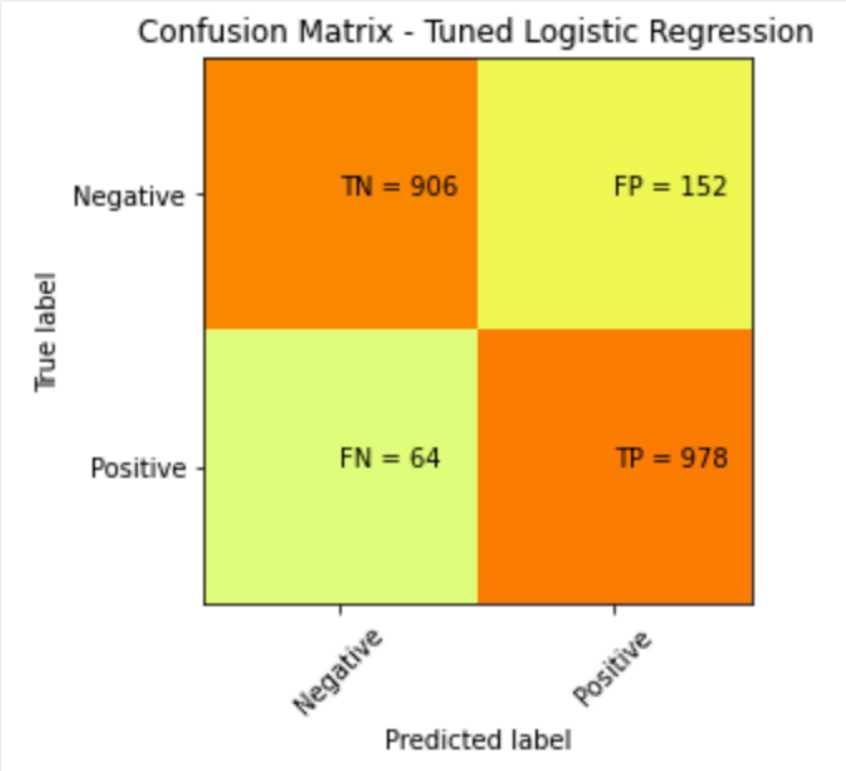
1. The initial dataset (3.5k positive and negative reviews)
2. The full dataset (12.5k positive and negative reviews)

Baseline Model – Tuned Logistic Regression

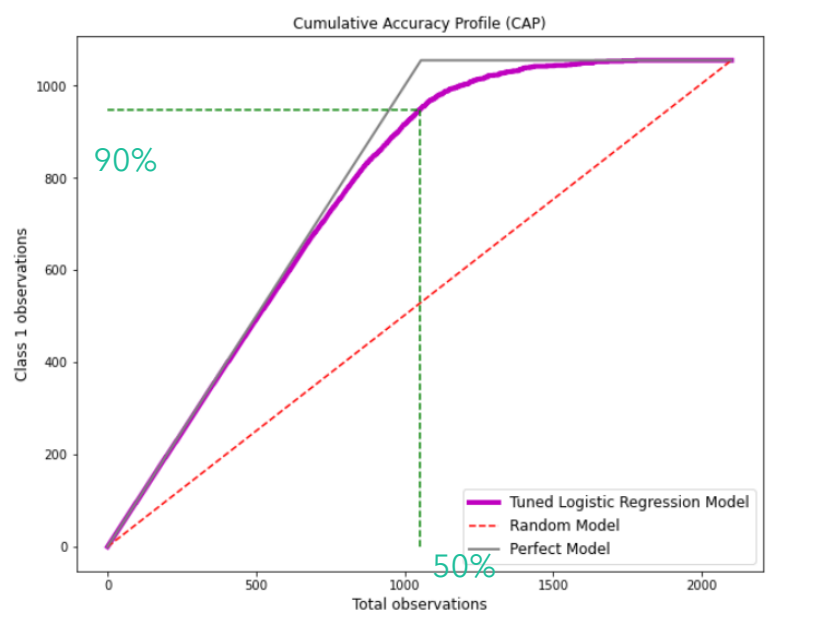
+ Following the model evaluation and ensemble learning analysis conducted in Milestone 3, the tuned logistic regression model was deemed the best based on accuracy and confusion matrix values.

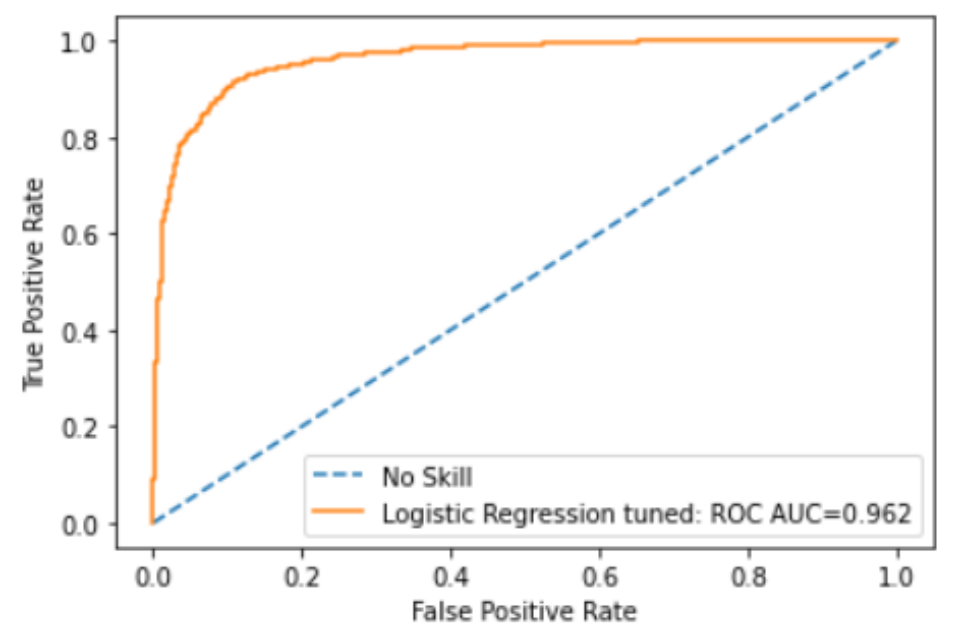
+ The tuned logistic regression model was our starting point for feature selection.

```
Model Performance metrics:
------------------------------
Accuracy: 0.8971
Precision: 0.9
Recall: 0.8971
F1 Score: 0.897

Model Classification report:
------------------------------
              precision    recall  f1-score   support

           1       0.87      0.94      0.90      1042
           0       0.93      0.86      0.89      1058

    accuracy                           0.90      2100
   macro avg       0.90      0.90      0.90      2100
weighted avg       0.90      0.90      0.90      2100
```



Tuned Logistic Regression





Guidelines for Model:

X > 90% Overfitting

80% < X < 90% Very Good Model

70% < X < 80% Good Model

60% < X < 70% Average Model

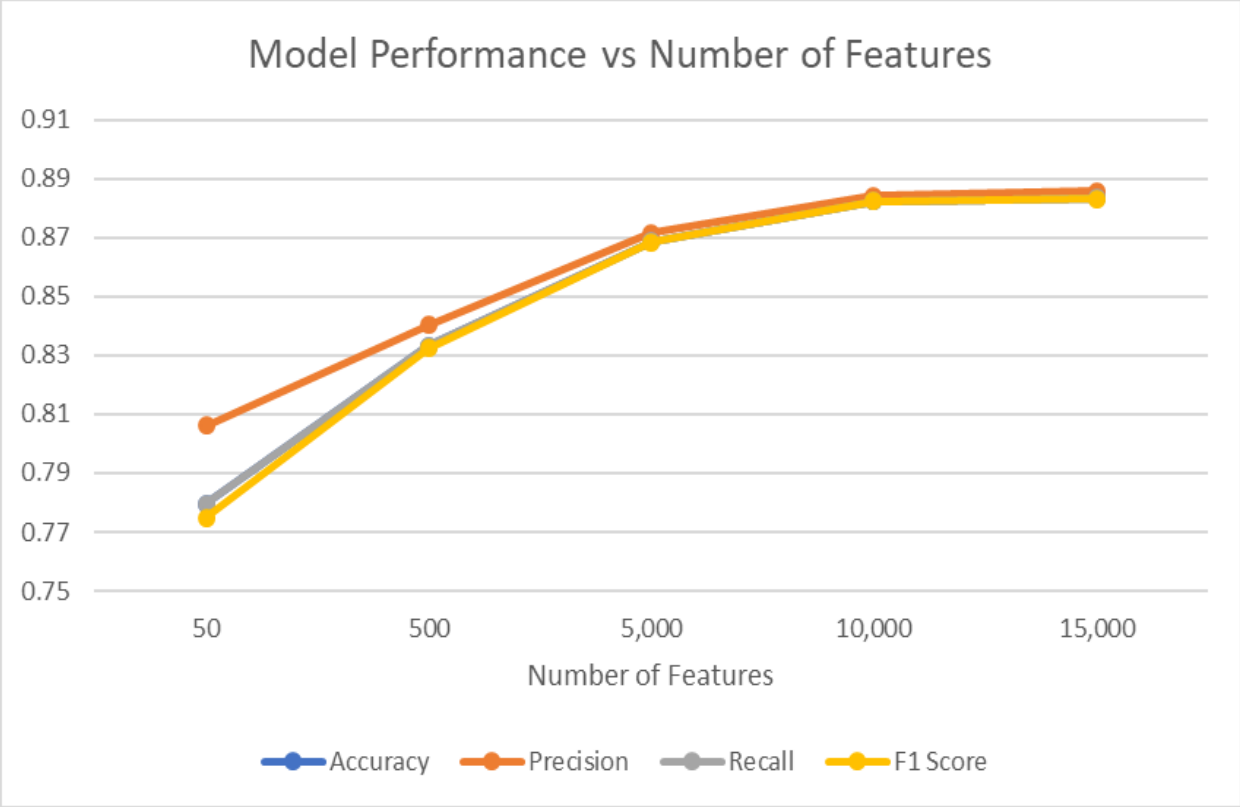
X < 60% Poor Model

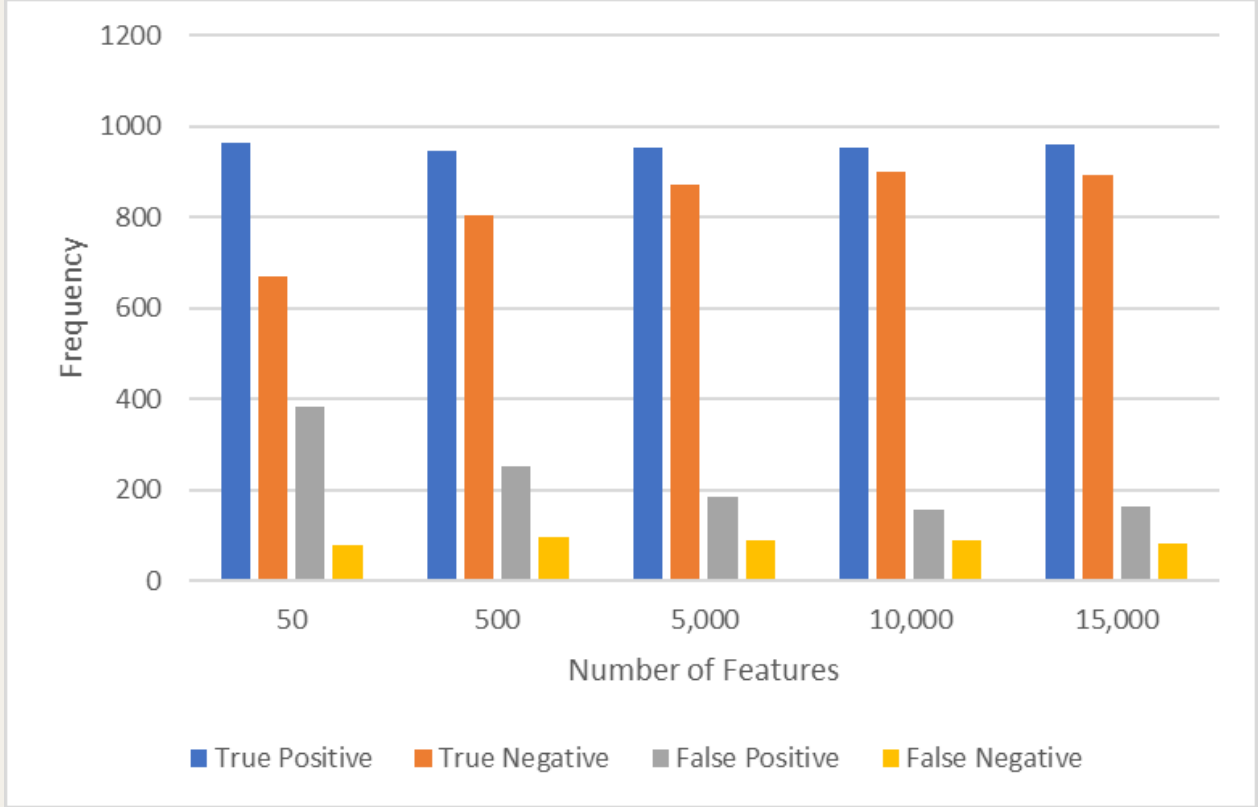
+ Accuracy ratio (AUTLR/AUP): 0.924
+ Evaluate the Model using the 50% line on the CAP curve:
  + Value at the 50% line is 90% which suggests that the tuned Logistic Regression model is very good

AUP = Area under the perfect model

AUTLR = Area under the tuned Logistic Regression model

Summary of Decision Tree Feature Selection





Slection Method	Features	Accuracy	Precision	Recall	F1 Score	CAP - Accuracy Rate	50th % line	TP	TN	FP	FN
Decision Tree	50	0.780	0.806	0.780	0.775	0.78	80.7%	965	672	385	78
Decision Tree	500	0.833	0.840	0.833	0.833	0.86	84.8%	945	805	252	98
Decision Tree	5,000	0.869	0.872	0.869	0.868	0.90	87.5%	952	872	185	91
Decision Tree	10,000	0.882	0.884	0.882	0.882	0.91	89.1%	954	899	158	89
Decision Tree	15,000	0.883	0.886	0.883	0.883	0.92	89.3%	961	894	163	82

Decision Tree Selection – Best Model (15k Features)

Guidelines for Model:

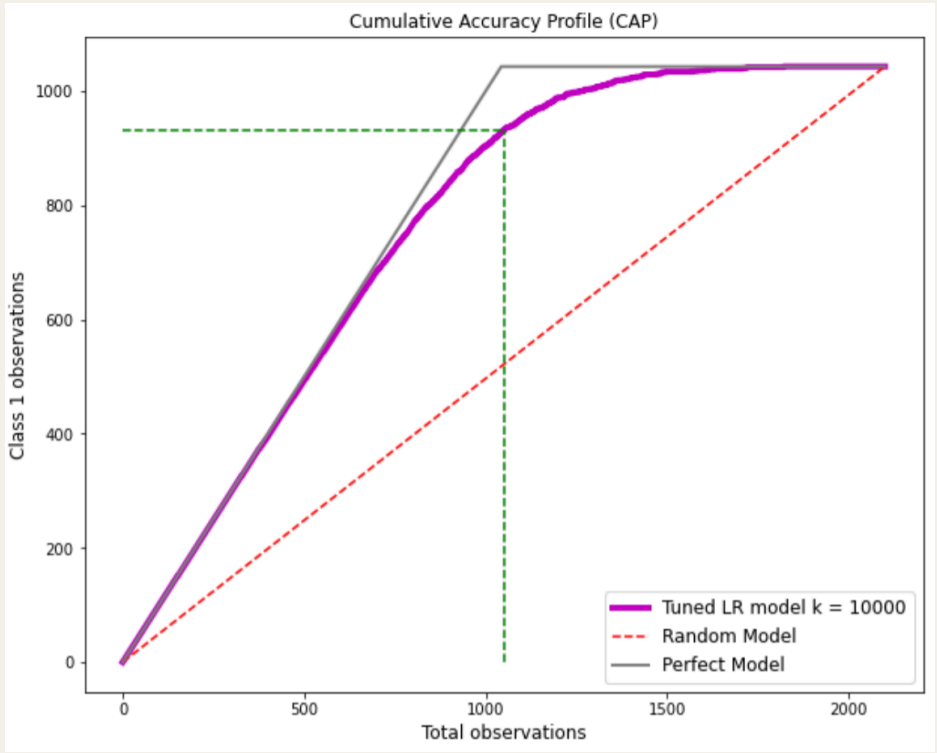
X > 90% Overfitting

80% < X < 90% Very Good Model

70% < X < 80% Good Model

60% < X < 70% Average Model

X < 60% Poor Model

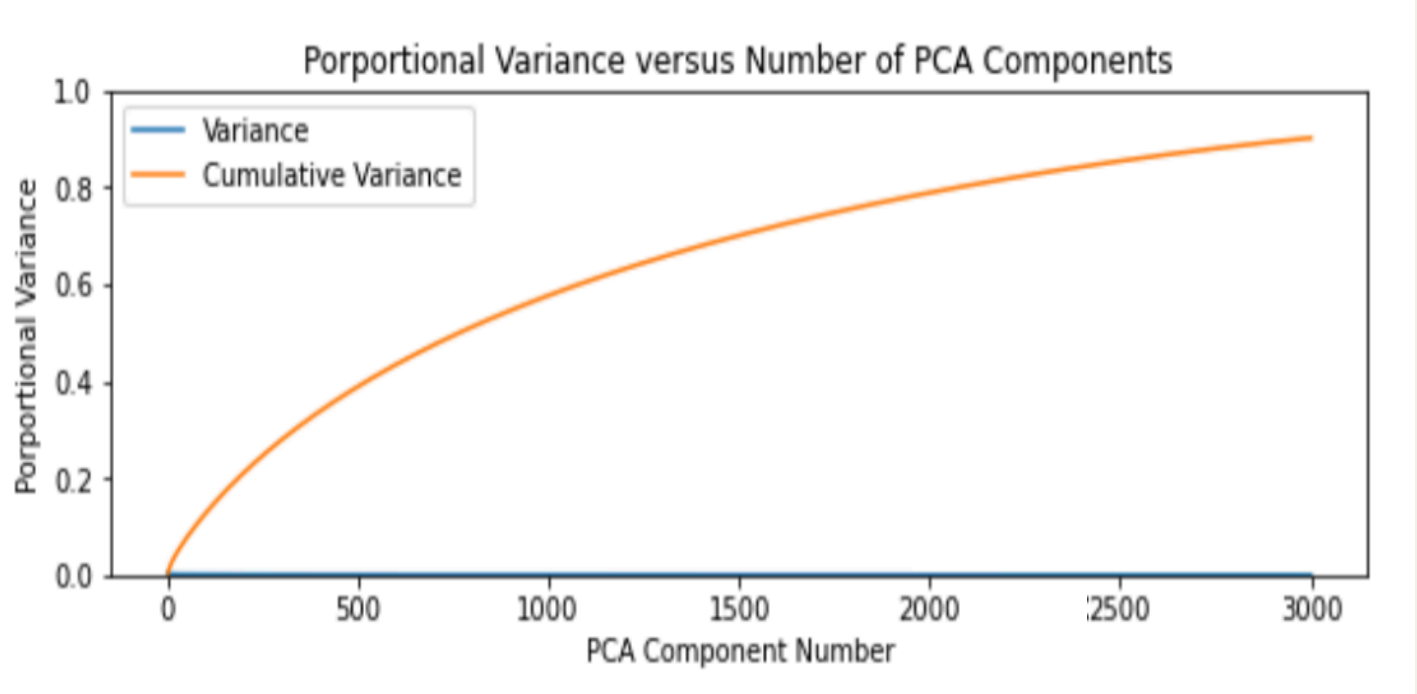


- Accuracy ratio (AUTLR/AUP): 0.924
- Evaluate the Model using the 50% line on the CAP curve:
  - Value at the 50% line is 89.26% which suggests that the tuned Logistic Regression model is a very good model

AUP = Area under the perfect model

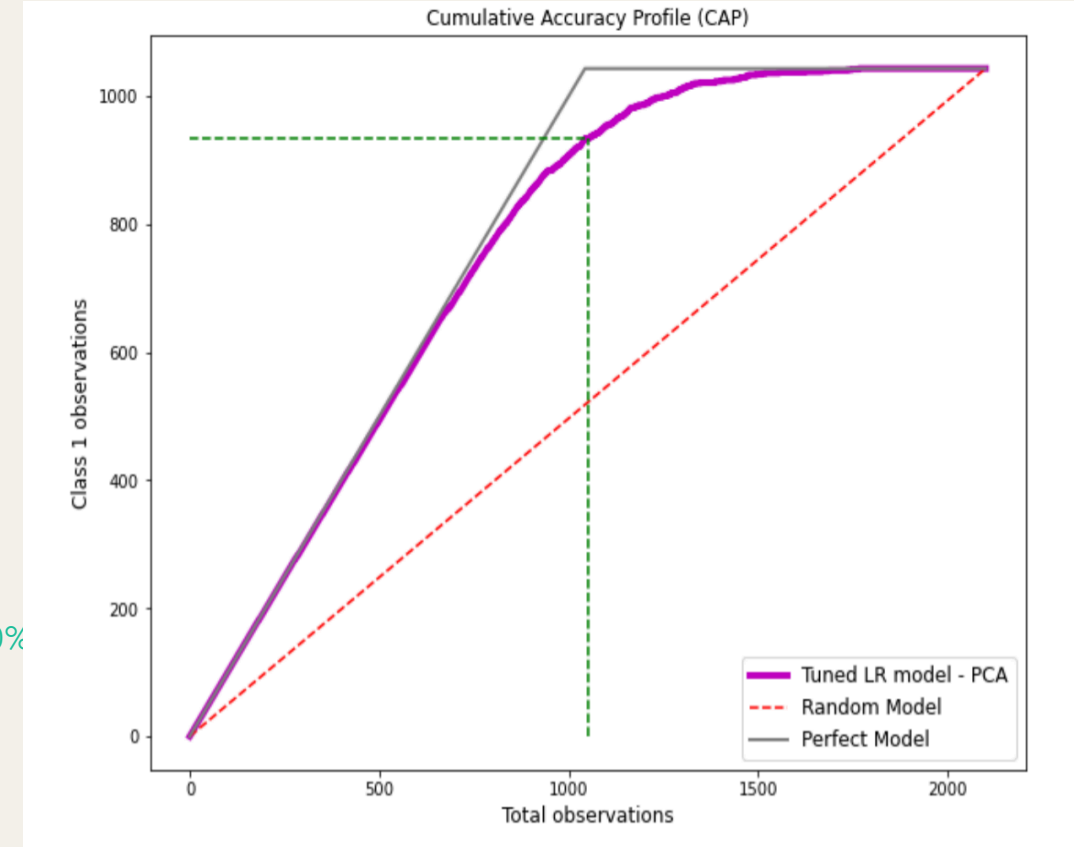
AUTLR = Area under the tuned Logistic Regression model

Principal Component Analysis (PCA)



Method	Components	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
PCA	3000	0.884	0.886	0.884	0.8837	0.923	89.55%	897	959	160	84

Tuned Logistic Regression PCA



50%

Guidelines for Model:

X > 90% Overfitting

80% < X < 90% Very Good Model

70% < X < 80% Good Model

60% < X < 70% Average Model

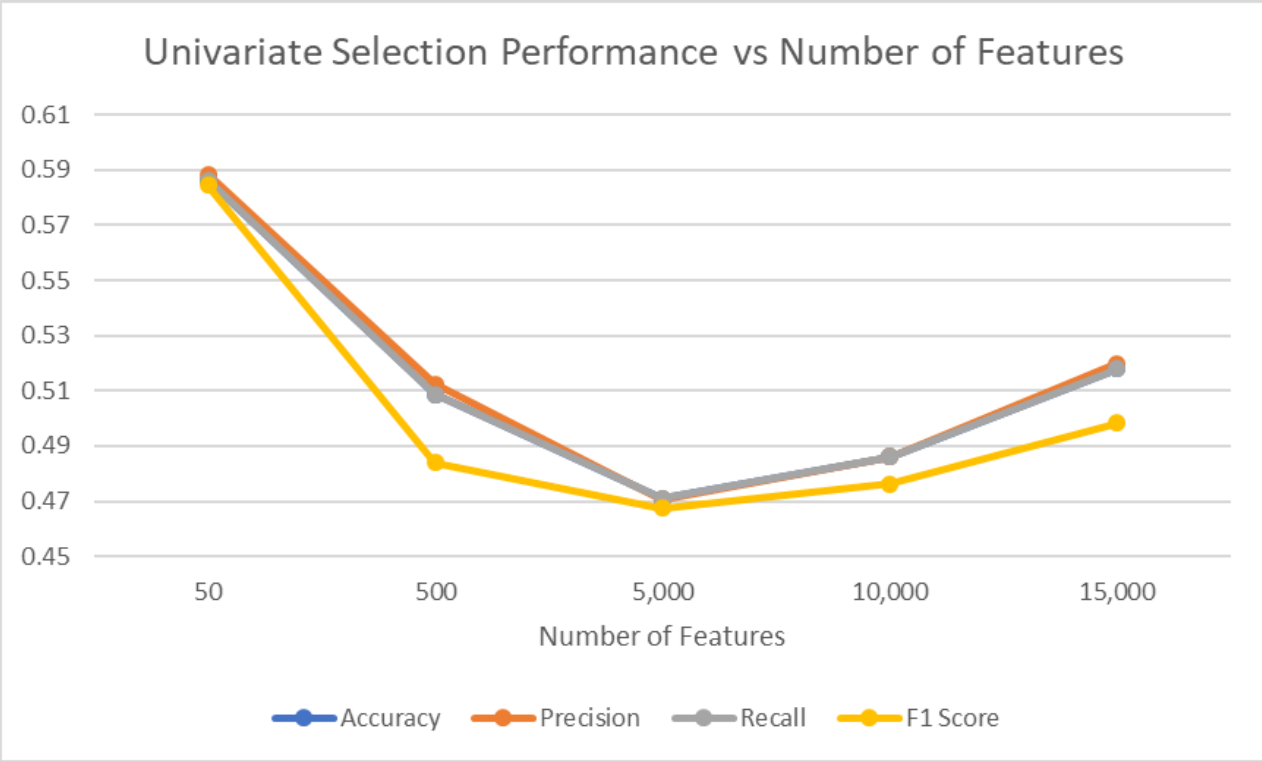
X < 60% Poor Model

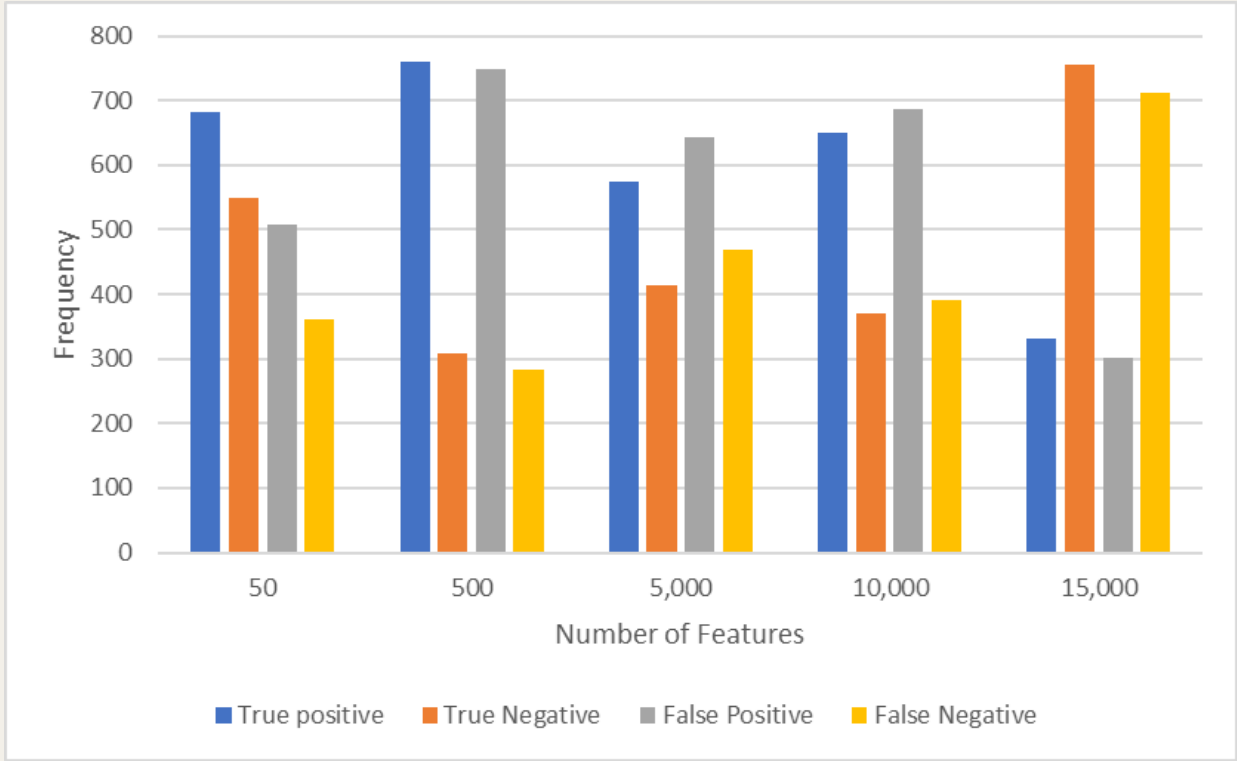
+ Accuracy ratio (AUTLR/AUP): 0.924
+ Evaluate the Model using the 50% line on the CAP curve:
  + Value at the 50% line is 89.549% which suggests that the tuned Logistic Regression model is a very good model

AUP = Area under the perfect model

AUTLR = Area under the tuned Logistic Regression model

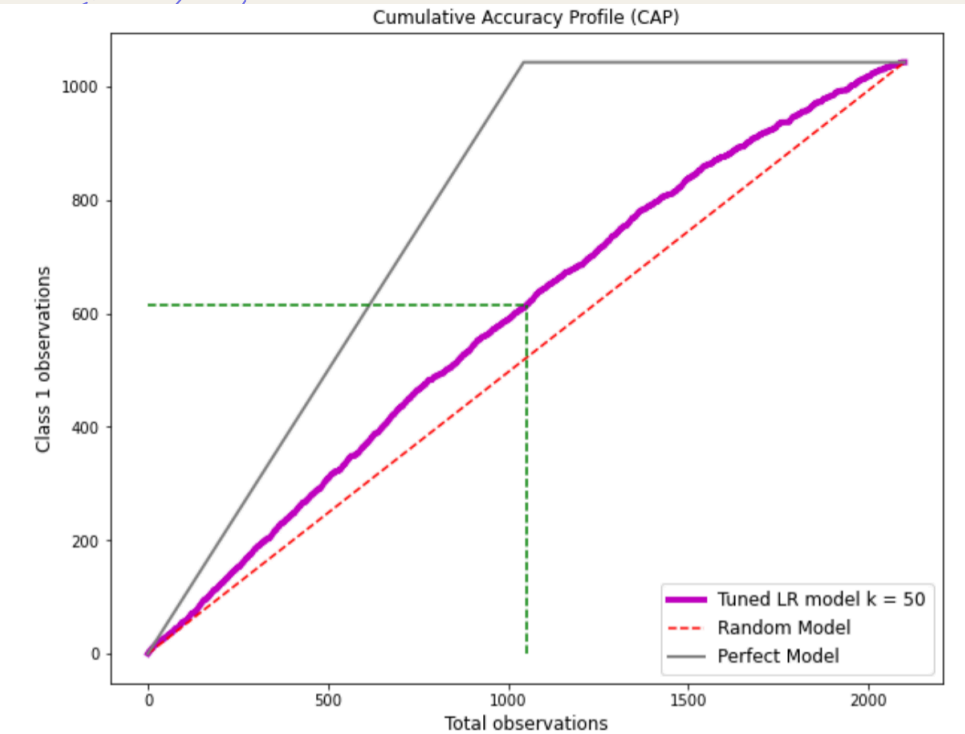
Summary of Univariate Feature Selection

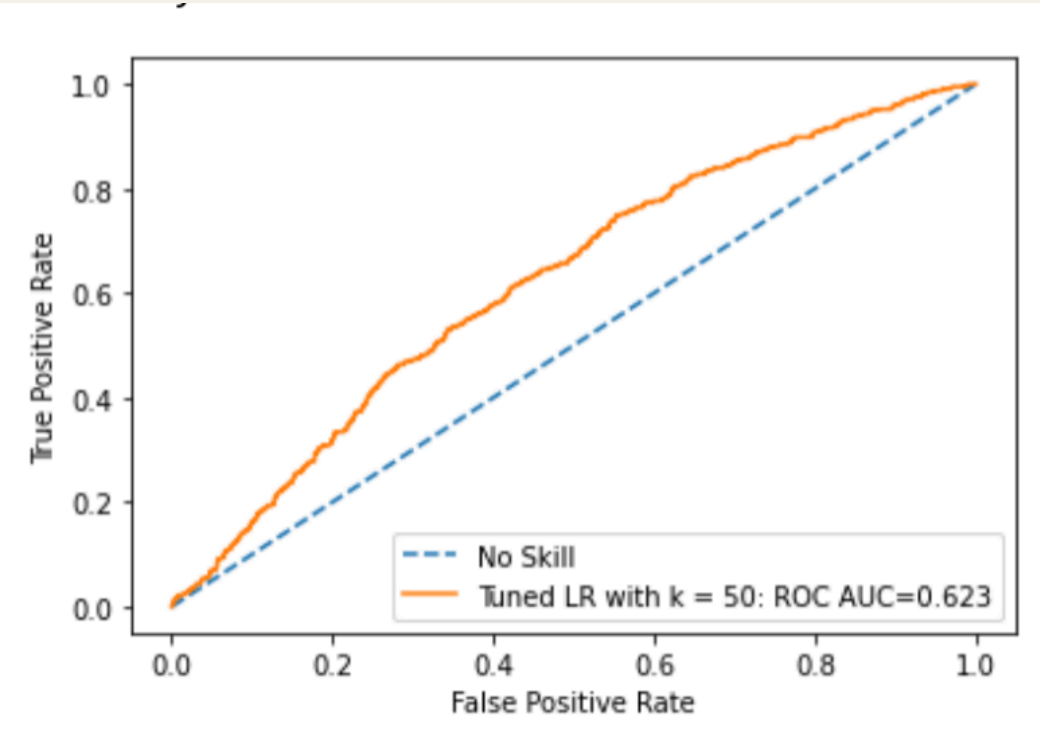




Slection Method	Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Univariate	50	0.586	0.588	0.586	0.584	0.245	58.9%	682	549	508	361
Univariate	500	0.509	0.513	0.509	0.484	0.006	50.0%	760	308	749	283
Univariate	5,000	0.471	0.471	0.471	0.468	-0.091	48.0%	575	414	643	468
Univariate	10,000	0.486	0.486	0.486	0.476	-0.062	48.0%	651	370	687	392
Univariate	15,000	0.518	0.520	0.518	0.498	0.051	51.0%	332	756	301	711

Univariate Feature Selection – Best Model (50 Feature)





Guidelines for Model:

X > 90% Overfitting

80% < X < 90% Very Good Model

70% < X < 80% Good Model

60% < X < 70% Average Model

X < 60% Poor Model

+ Accuracy ratio (AUTLR/AUP): 0.924
+ Evaluate the Model using the 50% line on the CAP curve:
  + Value at the 50% line is 58.896% which suggests that the tuned Logistic Regression model is a poor model

AUP = Area under the perfect model

AUTLR = Area under the tuned Logistic Regression model

Summary of Result – Initial Dataset

Model	Accuracy	Precision	Recall	F1 Score	TP	TN	FP	FN	ROC AUC	Accuracy Rate	50th % line
Baseline	0.897	0.9	0.897	0.897	978	906	152	64	0.965	0.930	90.3%

Slection Method	Features	Accuracy	Precision	Recall	F1 Score	TP	TN	FP	FN	Accuracy Rate	50th % line
Univariate	50	0.586	0.588	0.586	0.584	682	549	508	361	0.245	58.9%
Univariate	500	0.509	0.513	0.509	0.484	760	308	749	283	0.006	50.0%
Univariate	5,000	0.471	0.471	0.471	0.468	575	414	643	468	-0.091	48.0%
Univariate	10,000	0.486	0.486	0.486	0.476	651	370	687	392	-0.062	48.0%
Univariate	15,000	0.518	0.520	0.518	0.498	332	756	301	711	0.051	51.0%
Decision Tree	50	0.780	0.806	0.780	0.775	965	672	385	78	0.785	80.7%
Decision Tree	500	0.833	0.840	0.833	0.833	945	805	252	98	0.858	84.8%
Decision Tree	5,000	0.869	0.872	0.869	0.868	952	872	185	91	0.900	87.5%
Decision Tree	10,000	0.882	0.884	0.882	0.882	954	899	158	89	0.912	89.1%
Decision Tree	15,000	0.883	0.886	0.883	0.883	961	894	163	82	0.917	89.3%
PCA	3,000	0.884	0.886	0.884	0.884	897	959	160	84	0.923	89.5%

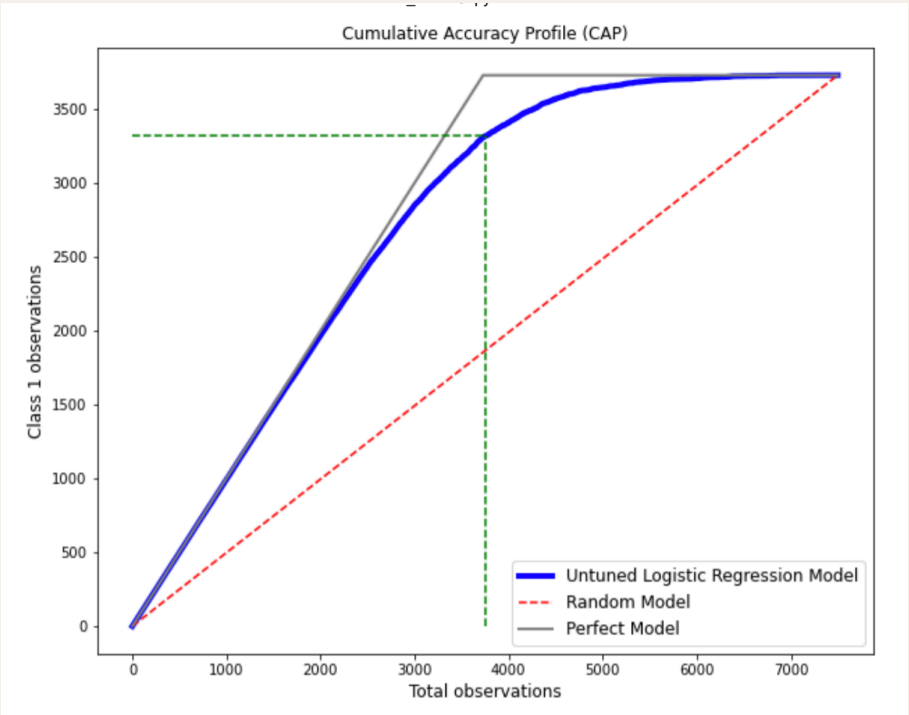
Improvement – Adding More data

- While trying to build upon the results obtained thus far and improve the model, our process was repeated using the full movie dataset
  - 25k observations
  - 12.5k positive and negative reviews
- The Logistic Regression model was baselined for the updated dataset:

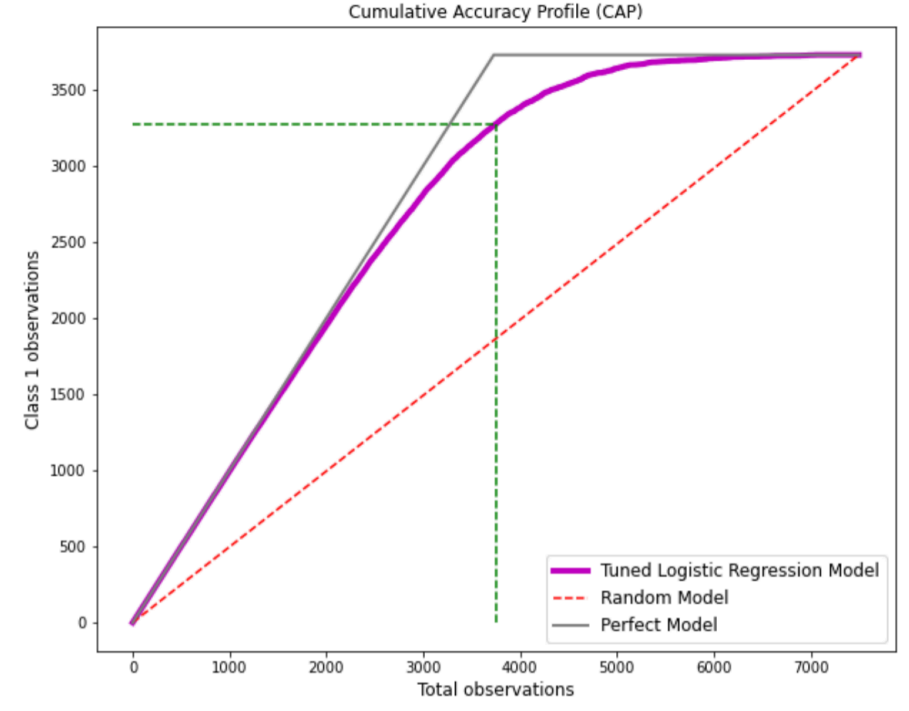
Model	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Baseline LR	0.884	0.884	0.884	0.884	0.910	88.9%	3349	3279	491	381
Tuned LR	0.872	0.874	0.872	0.872	0.898	87.8%	3382	3155	615	348

Logistic Regression Cap Curves

Baseline Logistic Regression



Tuned Logistic Regression



Guidelines for Model:

X > 90% Overfitting

80% < X < 90% Very Good Model

70% < X < 80% Good Model

60% < X < 70% Average Model

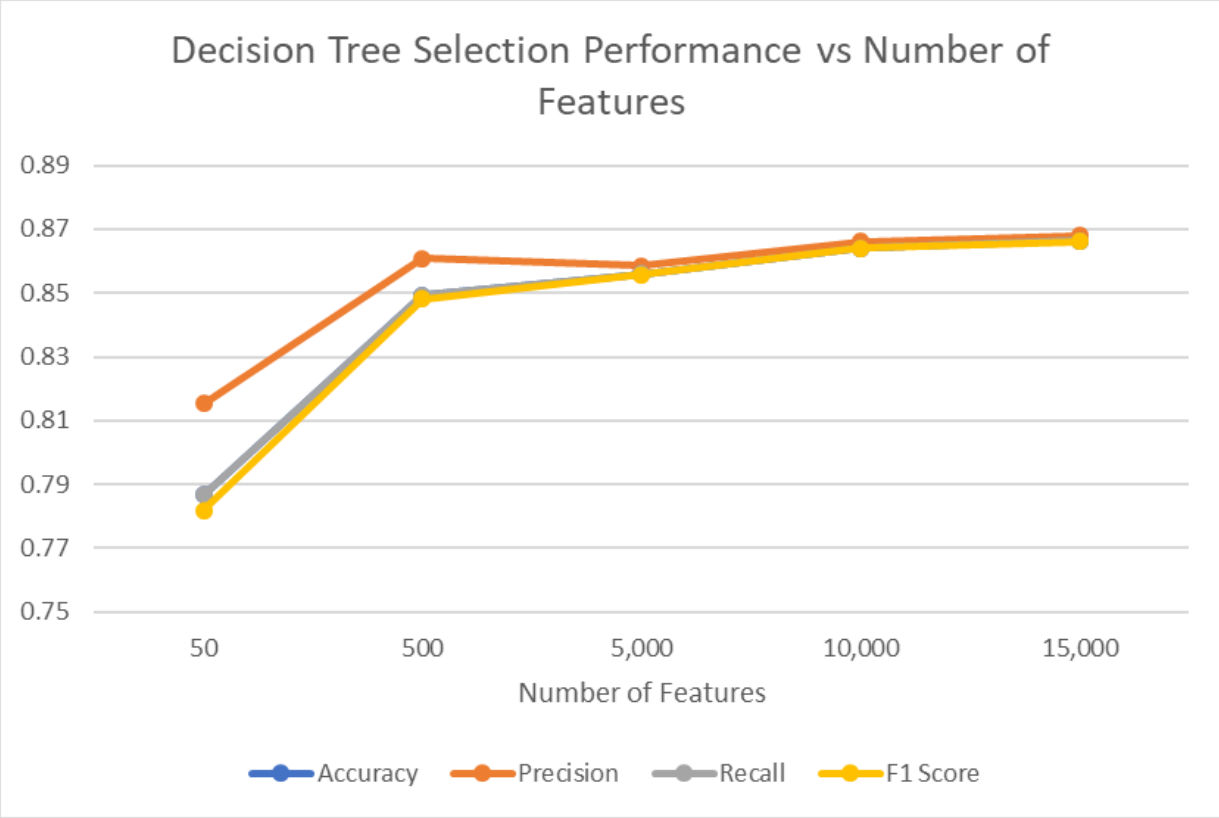
X < 60% Poor Model

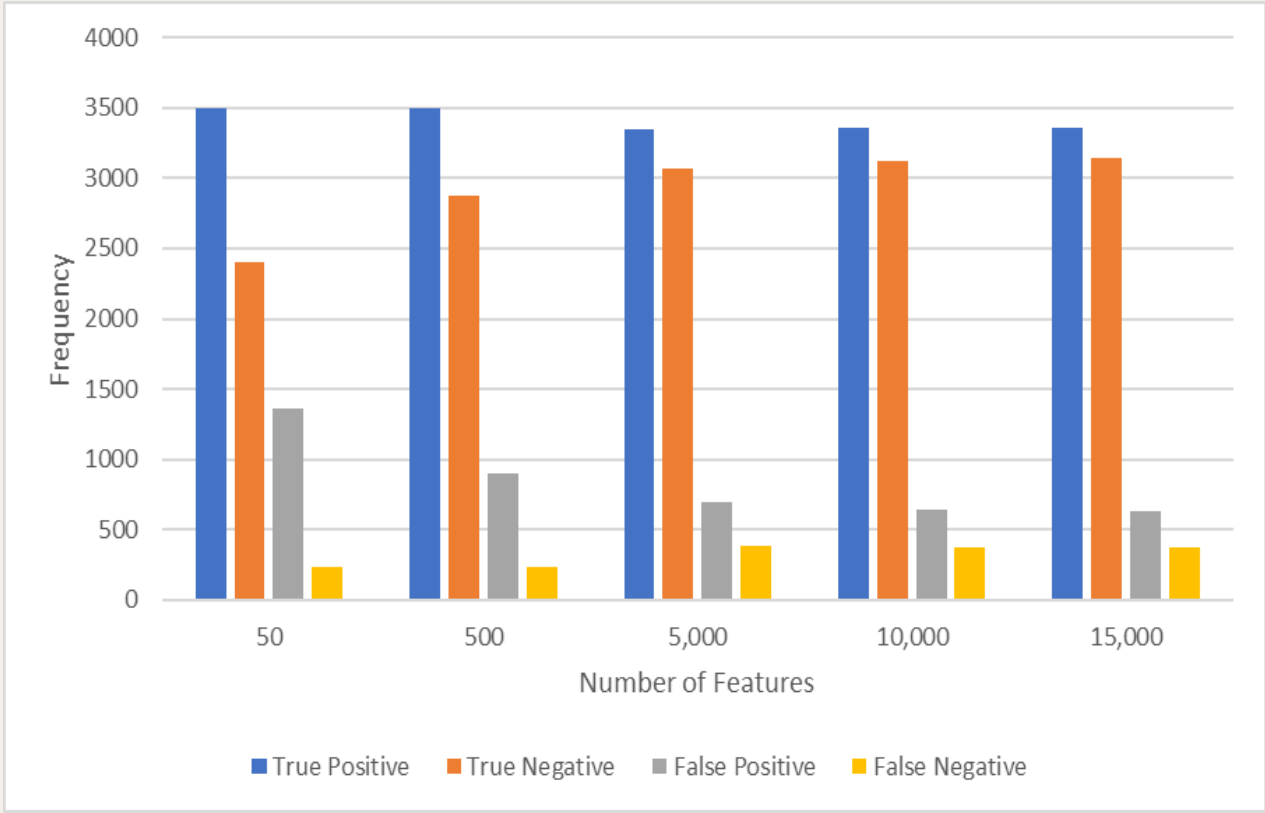
+ Accuracy ratio (AUTLR/AUP): 0.924
+ Evaluate the Model using the 50% line on the CAP curve:
  + Value at the 50% line is 88.92% for baseline and 97.77 % for tuned LR which suggest that the both Logistic Regression models are very good, with baseline being a slightly better model

AUP = Area under the perfect model

AUTLR = Area under the tuned Logistic Regression model

Summary of Decision Tree Feature Selection- Full Dataset

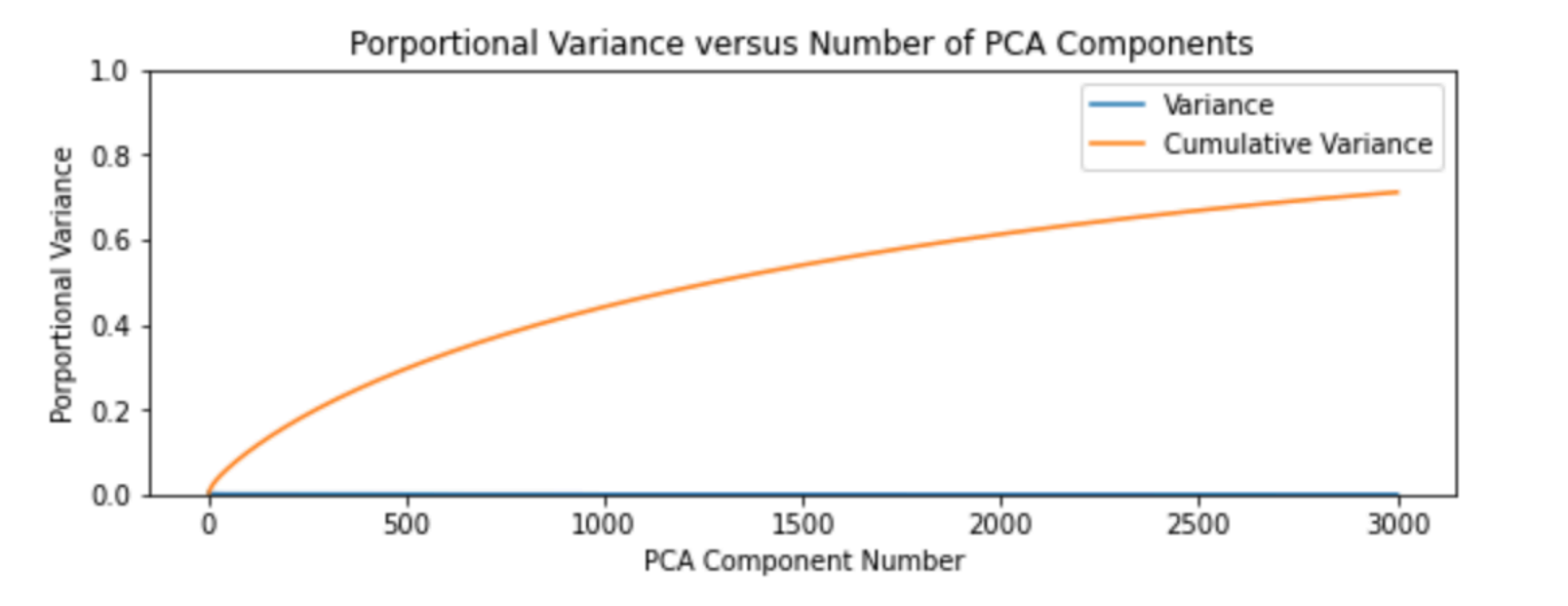




Slection Method	Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Decision Tree	50	0.787	0.815	0.787	0.782	0.790	81.4%	3493	2407	1363	237
Decision Tree	500	0.849	0.861	0.849	0.848	0.879	86.7%	3498	2872	898	232
Decision Tree	5,000	0.856	0.859	0.856	0.856	0.878	86.3%	3348	3072	698	382
Decision Tree	10,000	0.864	0.866	0.864	0.864	0.885	87.3%	3356	3125	645	374
Decision Tree	15,000	0.866	0.868	0.866	0.866	0.889	87.5%	3356	3141	629	374

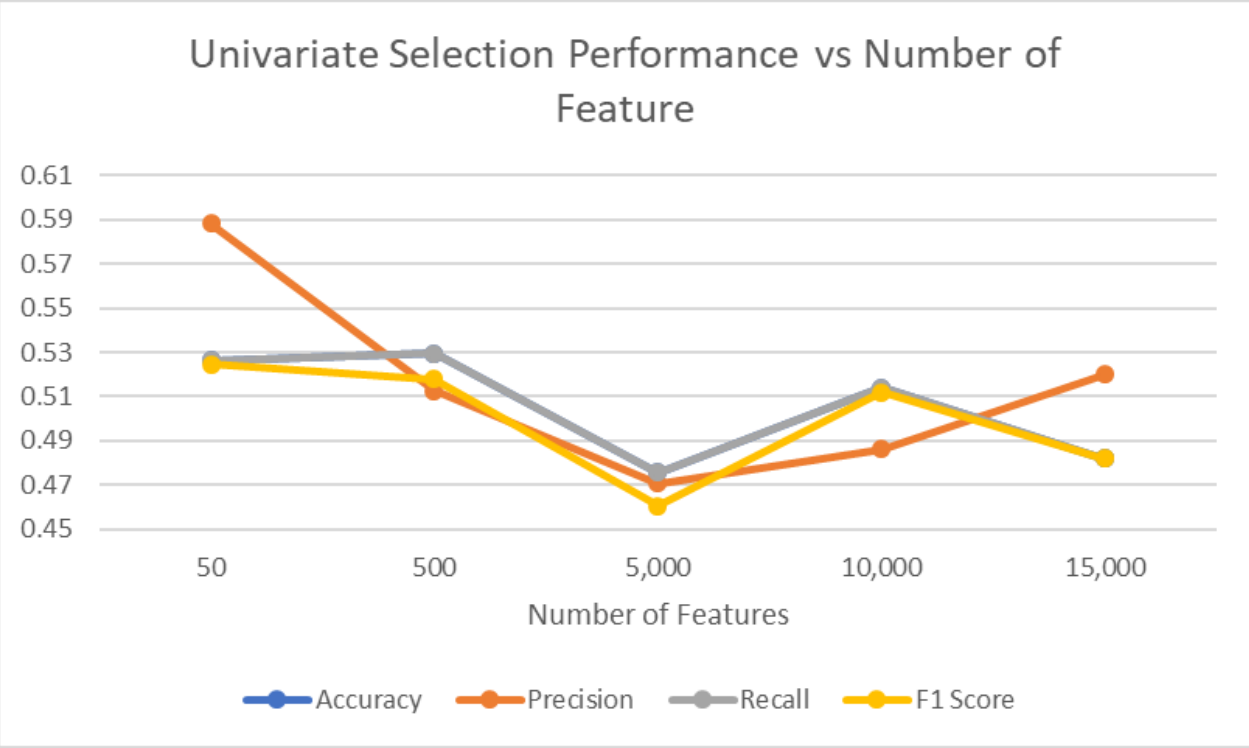
PCA

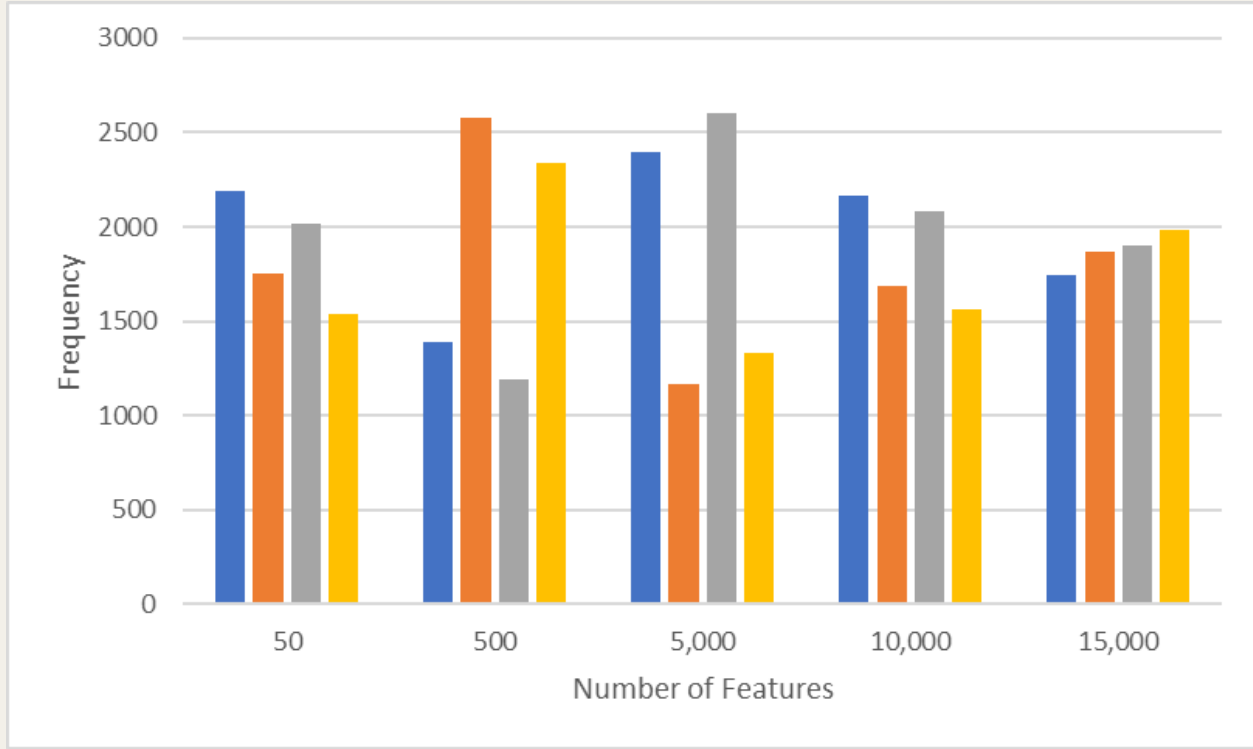
Selecting number of features (3000)



Method	Components	Accuracy	Precision	Recall	F1 Score	TP	TN	FP	FN	Accuracy Rate	50th % line
PCA	3000	0.868	0.871	0.868	0.868	3413	3095	675	317	0.895	87.6%

Summary of Univariate Feature Selection- Full Dataset





Slection Method	Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Univariate	50	0.526	0.527	0.526	0.525	0.041	52.4%	2194	1753	2017	1536
Univariate	500	0.530	0.532	0.530	0.518	0.106	53.5%	1392	2579	1191	2338
Univariate	5,000	0.476	0.473	0.476	0.461	-0.071	47.3%	2398	1168	2602	1332
Univariate	10,000	0.514	0.515	0.514	0.512	0.051	51.6%	2164	1690	2080	1566
Univariate	15,000	0.482	0.482	0.482	0.482	-0.053	48.1%	1746	1869	1901	1984

Summary of Results – Full Dataset

Model	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Baseline LR	0.884	0.884	0.884	0.884	0.910	88.9%	3349	3279	491	381
Tuned LR	0.872	0.874	0.872	0.872	0.898	87.8%	3382	3155	615	348

Selection Method	Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN
Univariate	50	0.526	0.527	0.526	0.525	0.041	52.4%	2194	1753	2017	1536
Univariate	500	0.530	0.532	0.530	0.518	0.106	53.5%	1392	2579	1191	2338
Univariate	5,000	0.476	0.473	0.476	0.461	-0.071	47.3%	2398	1168	2602	1332
Univariate	10,000	0.514	0.515	0.514	0.512	0.051	51.6%	2164	1690	2080	1566
Univariate	15,000	0.482	0.482	0.482	0.482	-0.053	48.1%	1746	1869	1901	1984
Decision Tree	50	0.787	0.815	0.787	0.782	0.790	81.4%	3493	2407	1363	237
Decision Tree	500	0.849	0.861	0.849	0.848	0.879	86.7%	3498	2872	898	232
Decision Tree	5,000	0.856	0.859	0.856	0.856	0.878	86.3%	3348	3072	698	382
Decision Tree	10,000	0.864	0.866	0.864	0.864	0.885	87.3%	3356	3125	645	374
Decision Tree	15,000	0.866	0.868	0.866	0.866	0.889	87.5%	3356	3141	629	374
PCA	3,000	0.868	0.871	0.868	0.868	0.895	87.6%	3413	3095	675	317

Analysis of Best Models

Using a subset of dataset (7000 rows)

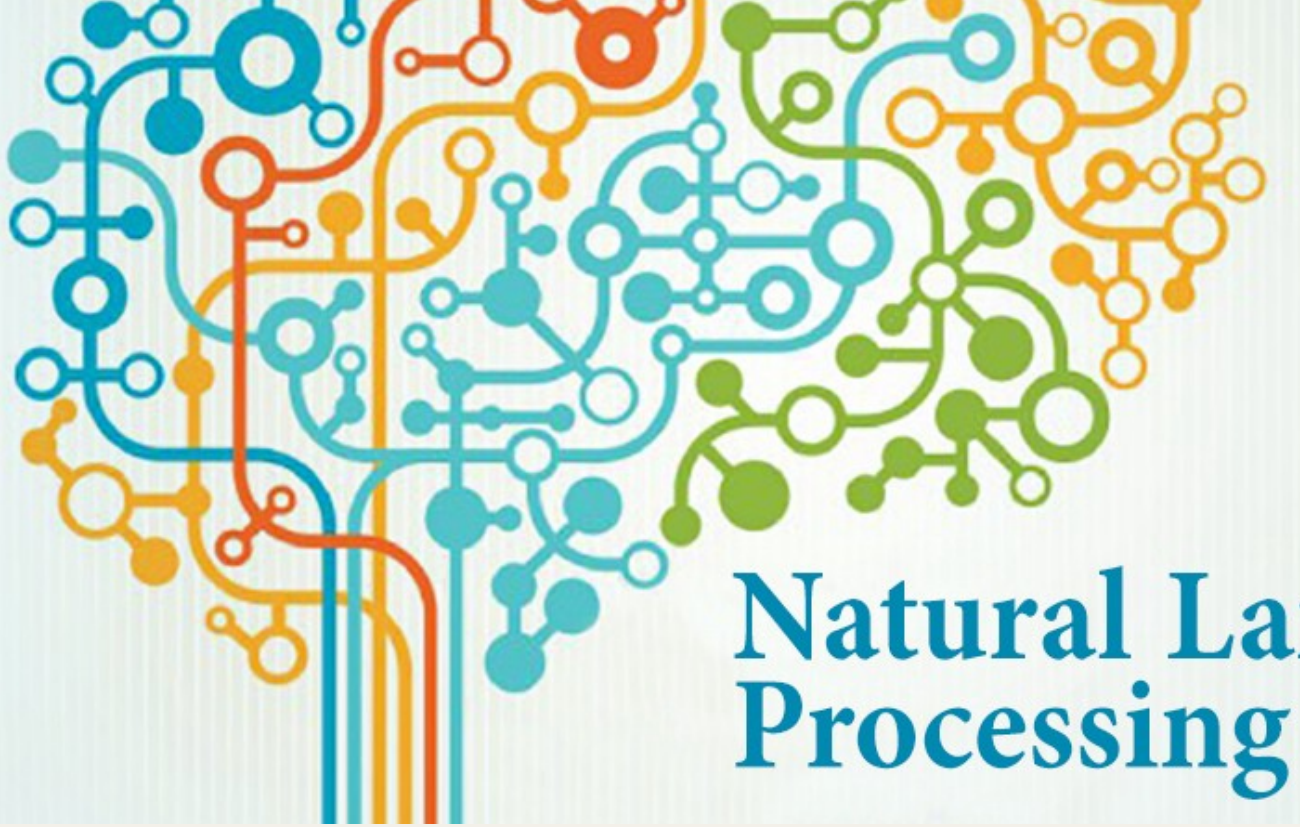
Method	Components/Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN	TP%	TN%	FP%	FN%
PCA	3000	0.884	0.886	0.884	0.884	0.923	89.5%	897	959	160	84	42.7%	45.7%	7.6%	4.0%
Decision Tree	15000	0.883	0.886	0.883	0.883	0.917	89.3%	961	894	163	82	45.8%	42.6%	7.8%	3.9%
Tuned LR	25000	**0.897**	0.900	0.897	0.897	0.930	90.3%	978	906	152	64	46.6%	43.1%	7.2%	**3.0%**

Using the full dataset

Method	Components/Features	Accuracy	Precision	Recall	F1 Score	Accuracy Rate	50th % line	TP	TN	FP	FN	TP%	TN%	FP%	FN%
PCA	3000	0.868	0.871	0.868	0.868	0.895	87.6%	3413	3095	675	317	45.5%	41.3%	9.0%	4.2%
Decision Tree	15000	0.866	0.868	0.866	0.866	0.889	87.5%	3356	3141	629	374	44.7%	41.9%	8.4%	5.0%
Baseline LR	25000	0.884	0.884	0.884	0.884	0.910	88.9%	3349	3279	491	381	44.7%	43.7%	6.5%	5.1%

Closing Remarks and Conclusions

- The main objective of this project was to develop a model which could take written movie reviews and classify their sentiment as either positive or negative.
- Following the machine learning life cycle process, the primary objective of this project was satisfied, and the following model was selected based on accuracy and confusion matrix analysis:
  - Tuned logistic regression model using the partial dataset on the full number of features extracted using BOW & TFIDF
- Interesting facts:
  - The data was obtained from standford.edu where the authors produced a paper analyzing the dataset. Though they used a different approach, they obtained an accuracy of 0.89
    - Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*



Thank You
Questions?

https://towardsdatascience.com/nlp-sequence-to-sequence-networks-part-2-seq2seq-model-encoderdecoder-model-6c22e29fd7e1