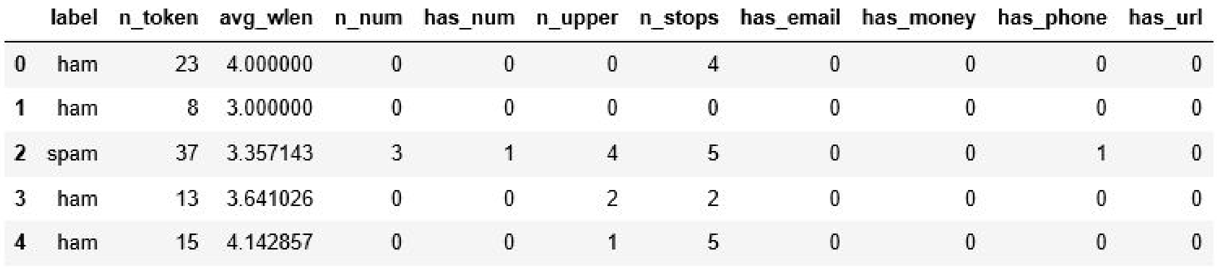
Capstone Project 1: Machine Learning Indepth-Analysis
Spring Board Data Science Career Track
~ By Zach Nguyen

The objective of the analysis is to build the machine learning part of a spam-filter. Thus, it is necessary to find the best way to predict the label of a given messahe with high precision and accuracy. After EDA, it is clear that we should explore in-depth machine learning analysis on the classical models as well as the bag of word models. Two dataframes are provided as input of the classical and the bag of word model. The dataframe with extracted features will be used to feed into Logistic Regression, Support Vector Classification, Random Forest and Gradient Boosted Trees. The raw text dataframe will be used to feed into Bi-gram models (Multinomial Naive Bayes, Support Vector Classification) with both Bag of Words and TF-IDF vectorizer. We will apply Scaling, Hyper-Parameter Tuning and GridSearch Cross validation techniques across all models to find the best fitting model.

A. Classical model

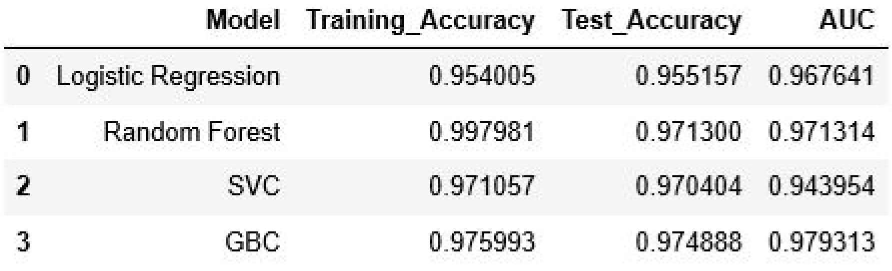
The final dataframe that will engender our feature space and label space can be seen as below:



	label	n_token	avg_wlen	n_num	has_num	n_upper	n_stops	has_email	has_money	has_phone	has_url
0	ham	23	4.000000	0	0	0	4	0	0	0	0
1	ham	8	3.000000	0	0	0	0	0	0	0	0
2	spam	37	3.357143	3	1	4	5	0	0	1	0
3	ham	13	3.641026	0	0	2	2	0	0	0	0
4	ham	15	4.142857	0	0	1	5	0	0	0	0

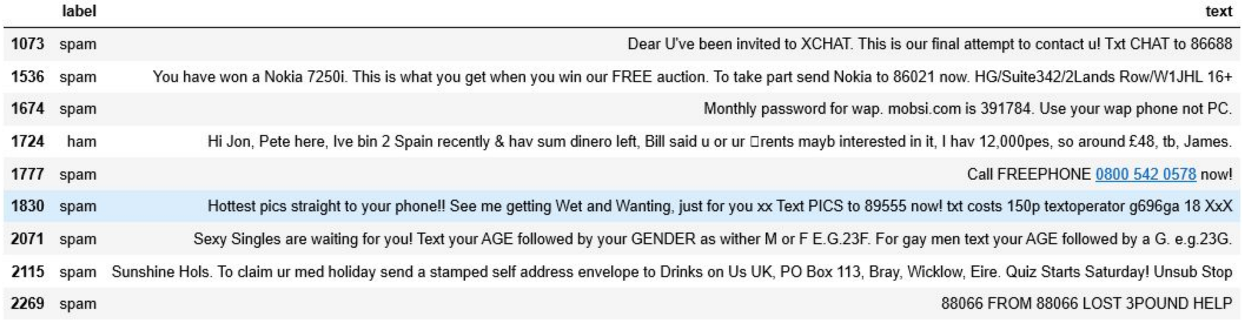
All the features were extracted manually, ready for scaling, divided into train and test set to be fed into our algorithms.

The final result of running the classical models can be seen from the table below:



	Model	Training_Accuracy	Test_Accuracy	AUC
0	Logistic Regression	0.954005	0.955157	0.967641
1	Random Forest	0.997981	0.971300	0.971314
2	SVC	0.971057	0.970404	0.943954
3	GBC	0.975993	0.974888	0.979313

Clearly, the king of the classical models is Gradient Boosting Tree, with 0.97 Test accuracy and 0.98 AUC score. This Ensemble Tree technique uses a learning mechanism to allow it to be punished for highly weighted misclassification and sample trees closer to accurate classification. Although crude, the classical model seems to be doing very well: having 97% accuracy pretty much has given you a decent professional spam filter. This baseline model looks hard to beat, but let's see some of its misclassifications:

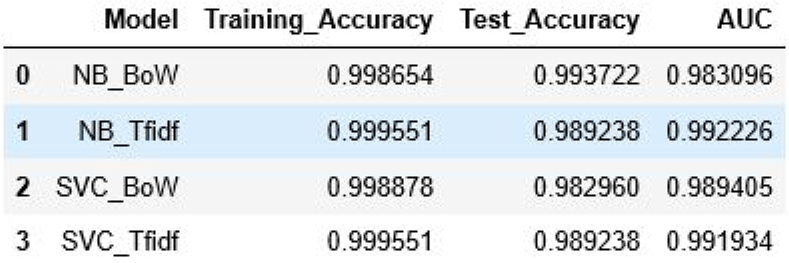


	label	text
1073	spam	Dear U've been invited to XCHAT. This is our final attempt to contact u! Txt CHAT to 86688
1536	spam	You have won a Nokia 7250i. This is what you get when you win our FREE auction. To take part send Nokia to 86021 now. HG/Suite342/2Lands Row/W1JHL 16+
1674	spam	Monthly password for wap. mobsi.com is 391784. Use your wap phone not PC.
1724	ham	Hi Jon, Pete here, Ive bin 2 Spain recently & hav sum dinero left, Bill said u or ur □rents mayb interested in it, I hav 12,000pes, so around £48, tb, James.
1777	spam	Call FREEPHONE 0800 542 0578 now!
1830	spam	Hottest pics straight to your phone!! See me getting Wet and Wanting, just for you xx Text PICS to 89555 now! txt costs 150p textoperator g696ga 18 XxX
2071	spam	Sexy Singles are waiting for you! Text your AGE followed by your GENDER as wither M or F E.G.23F. For gay men text your AGE followed by a G. e.g.23G.
2115	spam	Sunshine Hols. To claim ur med holiday send a stamped self address envelope to Drinks on Us UK, PO Box 113, Bray, Wicklow, Eire. Quiz Starts Saturday! Unsub Stop
2269	spam	88066 FROM 88066 LOST 3POUND HELP

Here we can see the drawback of this approach. SMSs with clearly flaggable words like 'Txt', 'Free', the punctuation '!' were ignored as a spam signal (a few urls, phone numbers were missed because of its uniqueness). We can further code these features into our existing model to improve it, but this clearly shows the advantage of the Bag of Word model we will explore next.

B. Bag of word model

With Bag of word model, we transform the train sets into Bag of Words and TD-IDF representation through scikitlearn CountVectorizer() and TfidfVectorizer(). The result is a Sparse Matrix training set. This matrix is fed into the Naive Bayes and Support Vector algorithms. The final result can be shown below.



	Model	Training_Accuracy	Test_Accuracy	AUC
0	NB_BoW	0.998654	0.993722	0.983096
1	NB_Tfidf	0.999551	0.989238	0.992226
2	SVC_BoW	0.998878	0.982960	0.989405
3	SVC_Tfidf	0.999551	0.989238	0.991934

Bag of word models are clearly superior to the traditional model. Naive Bayes and SVC are performing on par with each other, up to the 99% test accuracy mark with the Tf-idf transformation of vector space. A look at where it fails (Support Vector Classification) shows that it clearly inherits the Bag of Word model and misclassified a few messages due to its over-reliant on spammy and hammy words. However, these are understandable limitations of the Bag of Word model as it is unable to rely on context or semantics to classify.

	label	text
1672	ham	Glad to see your reply.
179	ham	Text her. If she doesnt reply let me know so i can have her log in
2402	spam	Babe: U want me dont u baby! Im nasty and have a thing 4 filthyguys. Fancy a rude time with a sexy bitch. How about we go slo n hard! Txt XXX SLO(4msgs)
2663	spam	Hello darling how are you today? I would love to have a chat, why dont you tell me what you look like and what you are in to sexy?
2770	spam	Burger King - Wanna play footy at a top stadium? Get 2 Burger King before 1st Sept and go Large or Super with Coca-Cola and walk out a winner
3360	spam	Sorry I missed your call let's talk when you have the time. I'm on 07090201529
731	spam	Email AlertFrom: Jeri StewartSize: 2KBSubject: Low-cost prescripiton drvgsTo listen to email call 123
751	spam	Do you realize that in about 40 years, we'll have thousands of old ladies running around with tattoos?