Gender Recognition by Voice

Farnoush Attarzadeh, Seyed Mohsen Hosseini, and Shaela Khan

Western University
1151 Richmond Street
London, Ontario, Canada
fattarza@uwo.ca
shosse59@uwo.ca
skhan923@uwo.ca

Abstract. Background: This era has seen massive improvements in the performance of artificial intelligence modelling algorithms to solve the task of predicting various features. And these techniques have shown great promise in various of scientific research. To that end- we have performed an experimentation on human vocal frequency and it's various attributes to predict whether the said voice is generated by a male or female. We have used logistic regression, PCA, Kmeans clustering, and Neural networks to build classifiers that performs that task of predicting the criteria. Multiple loss metrics and optimizers were used and discarded later on due to lack of performance improvement. We arrived at the conclusion that, there are certain features or attributes (if preferred)- of the human voice frequency which contribute greatly in distinguishing a male from a female vocal signature. The use and implications of these findings are open ended and far reaching.
Conclusion: We created classifiers that have decent accuracy and specificity but low sensitivity for predicting when those certain features were taken out of the training models. And those telling features were –(Q25, IQR, meanfun, label, sd, sp.ent, meanfreq, median, sfm, mode, centroid). With meanfun to be the ultimate feature for gender selection by voice signature. And Logistic regression to be a clear winner in performing the task predictably and with greater stability.

Keywords: Gender recognition · Kmeans · PCA · Logistic regression · Neural network

1 Introduction

We took it upon ourselves to solve the problem of prediction of gender based on voice signatures and it's various frequency properties as our guide. Voice recognition technology is a marvel of the latest advances of deep learning research, along with computer vision –which works to improve and essentially provide a computing machine with the ability to see and conceptualize; what exactly it is seeing very much like human beings do everyday. Machine learning advances-aid in the advances of these fields of research and application. And we are seeing

breathtaking results everyday. Voice recognition technology exploration is an active field of research with the advent of many voice command operated applications on our mobile phones, smart home companions, pc and various others. So, why is gender recognition an issue that needs to be investigated and tackled? The field of cyber security, recognizing a human being on a audio recording -which has multiple applications –are just some of the reason why- this problem needs a clever and reliable solution.

Recognizing the gender of a person based on their voice is not an easy and intuitive task for computing machine unlike human beings. In this paper therefore, we analyze whether a computing agent could perform such a task given proper training and resources -or not.

In order perform that task, we used a dataset retrieved from data.world[1] [1,2].This dataset has 21 columns and 3168 rows (1584 male and 1584 female). In table 1 a summery of the features in this dataset is presented. The next step is to find whether there is any correlation between the features of our dataset or not. There are multiple types of correlation metrics and we opted to use Pearson for our dataset.

1.1 Pearson Correlation

The Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 shows that as the value of one variable increases, so does the value of the other variable. A value less than 0 shows that as the value of one variable increases, the value of the other variable decreases. in figure 1 the correlation between columns is presented.

As reported by the figure 1, there is a high correlation between (label, meanfun), (label, IQR), (label, sd), (label, Q25), (label, sp.ent). Nevertheless, we all now that correlation is not causation. What we could do is train our dataset solely by using each of the features and then compare the results. In the table 2 you can see the results of training on the dataset using logistic regression.

The results indicate that, if we include only one single feature 'meanfun' – we could possibly predict with 95% accuracy. And if we made use of -all the features for making predictions– our models reach an accuracy of 99% accuracy. Which is quite high , not to mention introduces a possibility of overfitting hence making the model less reliable in making predictions.

As it is shown by the figure 1, the meanfun has the most correlation with the label. Meanfun is the average of fundamental frequency measured across acoustic signal which is kind of an estimation of the wavelength of the first wave made in the pharynx. As the length of the pharynx in men is typically larger than women, it makes sense that the wavelengths of their standing waves be different and this variable be so important. But, why other variables are effective too?

[1] https://data.world/ml-research/gender-recognition-by-voice

Because signal of voice can be really different when people shout or sing or even talk differently, the frequency can be changed with how the pharynx is used and filled and emptied with air. Also there are a lot more effective parameters like how the signal is rerecorded, with what angle and etc [10].

Therefore, in order to introduce complexity and hone our combined creative problem-solving skills –we performed feature selection. We removed the following features from our dataset –(Q25, IQR, meanfun, label, sd, sp.ent, meanfreq, median, sfm, mode, centroid). With this change, the accuracy was around equal to 70% across all the methods we used to create our classifiers.

2 Data visualization and Clustering

In order to visualize this dataset we reduced it's dimensions to 2 or 3.PCA is one of the most well-known algorithms used for dimensionality reduction[8, Chapter 10.2]. Figure 2 shows the data in 2-dimensional space and figure 3 demonstrates it in 3-dimensional space.

As can be seen in both figure 2 and figure 3, there were no significant differences in terms of distance between male and female voices in this unique dataset. Hence, we had established that clustering would not be an ideal classifier in distinguishing male/female voice with high frequency.

There are various classes of clustering algorithms such as DBSCAN [7],Kmeans [9] and minibatch kmeans – which have been proven to be successful and efficient in predictive modelling tasks [11]. Since we already know the number of clusters and the dataset is not too big, we had used kmeans in this paper. And our previous assumptions were proven to be true -the accuracy of this method was 52%- which is slightly better than random.

3 Methods

All data pre-processing and analyses were done in python 3.7 [12]. The neural-network was designed and implemented with tensorflow 1.7 as back-end [5], keras [3] and sklearn [4] libraries for most modelling tasks

3.1 Logistic Regression

In this section we expound the logistic regression [8, Chapter 4.3] part of our project. In a nutshell, we tried different subsets of features to make the problem more challenging. We computed the confidence interval for each subset and plotted ROC and AUC diagrams in order to compare the results. It worth to be noted that in both phases we used 10-fold cross-validations [8, Chapter 5.1].

3.1.1 Phase I (hold-3-out Prediction)

As stated earlier, our focus is to predict the gender of a person based on the features presented in this dataset. In the first section we showed what will happen

if we only use a single feature to predict the gender. As presented in table 2, meanfun, IQR, and Q25, are the features that could predict the gender with a high level of accuracy. As a result, we thought that it could be more challenging if we dropped this three features from the dataset and determined its effect on the accuracy of our predictions.

We used both 'l1' and 'l2' as penalty of our logistic regression model. Nevertheless, there were not any big differences between the results. Therefore, we only report logistic regression with 'l1' penalty. Dropping these features have a big effect on the results because the accuracy is equal to 86%. 95.0 % of confidence interval is between 82.8% and 86.1%. The figure 4 shows the distribution of the accuracy by using bootstrapping with 300 different samples. Figure 5 shows the Area Under the Curve (AUC) which is the measure of the ability of a classifier to distinguish between classes. Figure 6 presents the Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems.

3.1.2 Phase II (hold-9-out Prediction)

Now let us remove all 9 important features from the dataset. With these changes, the accuracy would be equal to 70%. 95.0 % of confidence interval is between 69.4% and 73.1%. the figure 7 shows the distribution of the accuracy by using bootstrapping with 300 different sample. Figure 8 shows the Area Under the Curve (AUC) which is the measure of the ability of a classifier to distinguish between classes. Figure 9 presents the Receiver Operator Characteristic (ROC) curve which is an evaluation metric for binary classification problems.

3.2 Neural Network

For this part of the classifier, we used a simple multi-layer perceptron. Architecture of our method presented in figure 10.

We tried as Optimizer - rmsprop, sgd, Adam etc- with Adam performing best. BinaryCrossentropy to calculate loss metric since- the results are binary with 1 and 0 signifying male/female predictions. We used SparseCategoricalCrossEntropy as an experimental setup, which performed with 49% train accuracy and 37% test accuracy. Therefore, it was dropped as measure to evaluate the performance [6].

Epoch = 100
loss: 0.55 - accuracy: 0.68 **In test:** - Loss: 0.51 - accuracy: 0.67
Epoch = 500
loss: 0.49 - accuracy: 0.69 **In test:** - Loss: 0.51 - accuracy: 0.69
Epoch = 700
loss: 0.43 - accuracy: 0.69 **In test:** - Loss: 0.51 - accuracy: 0.69
Epoch = 1000
loss: 0.45 - accuracy: 0.77 **In test:** - Loss: 0.51 - accuracy: 0.77
Epoch = 10000
loss: 0.35 - accuracy: 0.77 **In test:** - Loss: 0.31 - accuracy: 0.77

4 Result

It was established early on that –certain features were more important contributors to perform a prediction reliably. Most of those features were taken out to perform these experiments. PCA and Kmeans classifier performed a little better than random guess-work. And Neural Networks Logistic Regression performed the best among all of the methods. In the figure 11 we compare our methods in different training size.

Linear regression capped out at - Accuracy: 0.702 Recall: 0.703 Precision: 0.722.

Neural Network capped out at – loss: 0.4496 - accuracy: 0.7680 – epoch =1000 For test data - Loss: 0.5135 - accuracy: 0.7666

Kmeans -accuracy: 0.52

5 Discussion

We developed predictive models to classify 1) gender based on some vocal signature data – which was curated from several hundred audio recordings.

We can theorize – given that we dropped the main predictors of voice that differentiates a male and female voice signature- the neural networks was less efficient in performance, however where Neural network classifier failed , the Logistic regression classifier triumphed.

What we can deduce from these findings – is something very interesting.

• There are certain features or attributes of voice frequencies which are strong predictors of male or female entity.

• Now that we are aware- we can simply use these features to make better, more robust technology for the hearing impaired.

• The performance of neural network though touted as a robust and super effective learning model – has flaws – where strong predictive features are left out of the training equation.

Overall, our findings were considered a success and a great learning exercise for future work in the field of data-science.

References

1. Actual dataset explanation and source vault (2020 (accessed December 3, 2020)), http://festvox.org/cmu_arctic/
2. dataset source and explanation (2020 (accessed December 3, 2020)), http://nsi.wegall.net/
3. Keras (2020 (accessed December 3, 2020)), https://keras.io/
4. SKlearn (2020 (accessed December 3, 2020)), [11]https://scikit-learn.org/stable/
5. TensorFlow (2020 (accessed December 3, 2020)), https://www.tensorflow.org/resources/tools
6. Aggarwal, C.C., et al.: Neural networks and deep learning. Springer (2018)

7. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
8. Gareth, J., Daniela, W., Trevor, H., Robert, T.: An introduction to statistical learning: with applications in R. Spinger (2013)
9. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. JSTOR: Applied Statistics **28**(1), 100–108 (1979)
10. Mongia, P.K., Sharma, R.: Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. Journal of Computer Networks and Communications **2014** (2014)
11. Sculley, D.: Web-scale k-means clustering. In: Proceedings of the 19th international conference on World wide web. pp. 1177–1178 (2010)
12. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)

Appendix

Table 1: dataset features

meanfreq	mean frequency (in kHz)	**sd**	standard deviation of frequency
median	median frequency (in kHz)	**Q25**	first quantile (in kHz)
Q75	third quantile (in kHz)	**IQR**	interquantile range (in kHz)
skew	skewness	**kurt**	kurtosis
sp.ent	spectral entropy	**sfm**	spectral flatness
mode	mode frequency	**centroid**	frequency centroid
meanfun	average of fundamental frequency measured across acoustic signal	**minfun**	minimum fundamental frequency measured across acoustic signal
maxfun	maximum fundamental frequency measured across acoustic signal	**meandom**	average of dominant frequency measured across acoustic signal
mindom	minimum of dominant frequency measured across acoustic signal	**maxdom**	maximum of dominant frequency measured across acoustic signal
dfrange	range of dominant frequency measured across acoustic signal	**modindx**	modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range

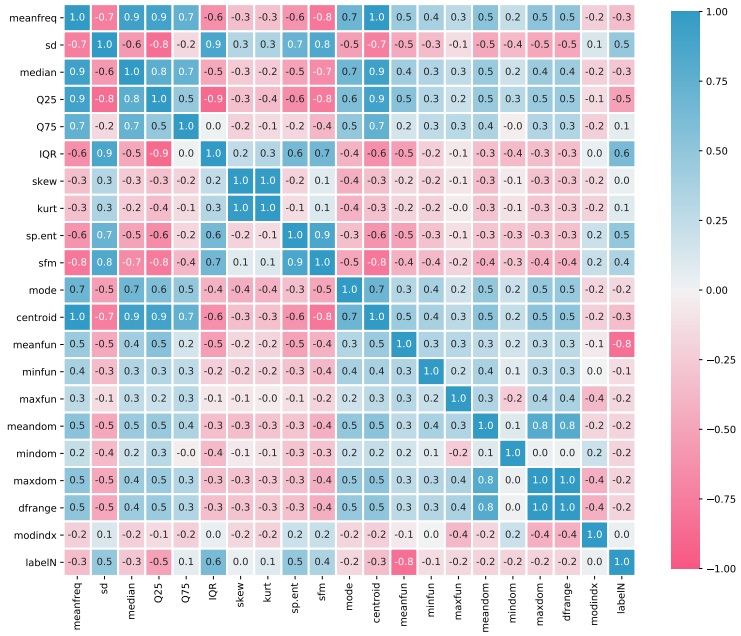

Fig. 1: Pearson correlation between columns in dataset.

Table 2: Dataset learning result.

feature	Accuracy	Recall	Precision	feature	Accuracy	Recall	Precision
meanfreq	0.65	0.57	0.71	sd	0.78	0.82	0.78
median	0.64	0.57	0.69	Q25	0.85	0.86	0.86
Q75	0.53	0.5	0.55	IQR	0.89	0.93	0.87
skew	0.49	0.08	0.57	kurt	0.49	0.08	0.62
sp.ent	0.74	0.73	0.76	sfm	0.65	0.59	0.69
mode	0.67	0.52	0.78	centroid	0.65	0.57	0.71
meanfun	0.95	0.97	0.95	minfun	0.53	0.5	0.56
maxfun	0.51	0.32	0.57	meandom	0.57	0.6	0.59
mindom	0.59	0.8	0.58	maxdom	0.57	0.55	0.6
dfrange	0.57	0.56	0.6	modindx	0.48	0.23	0.51

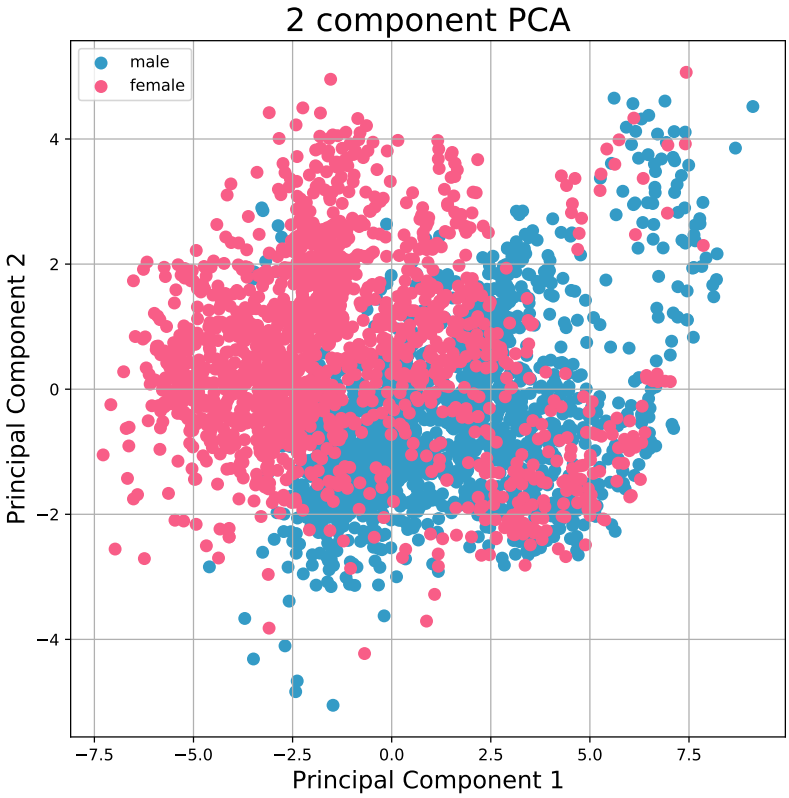

Fig. 2: 2d presentation of the dataset.

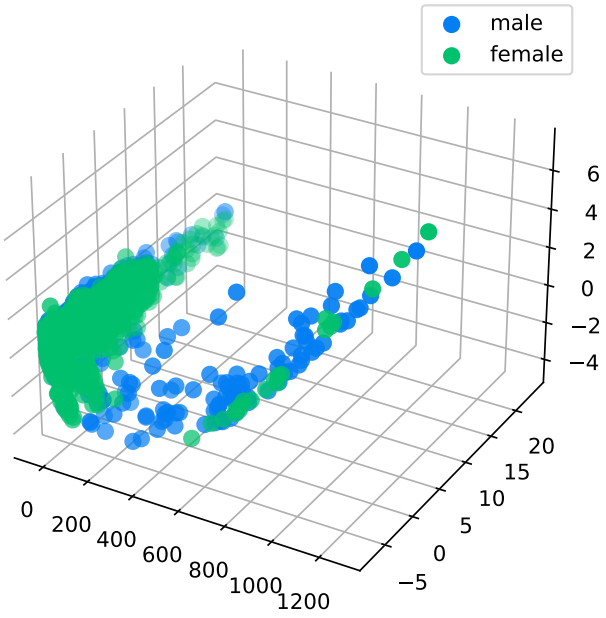

Fig. 3: 3d presentation of the dataset.

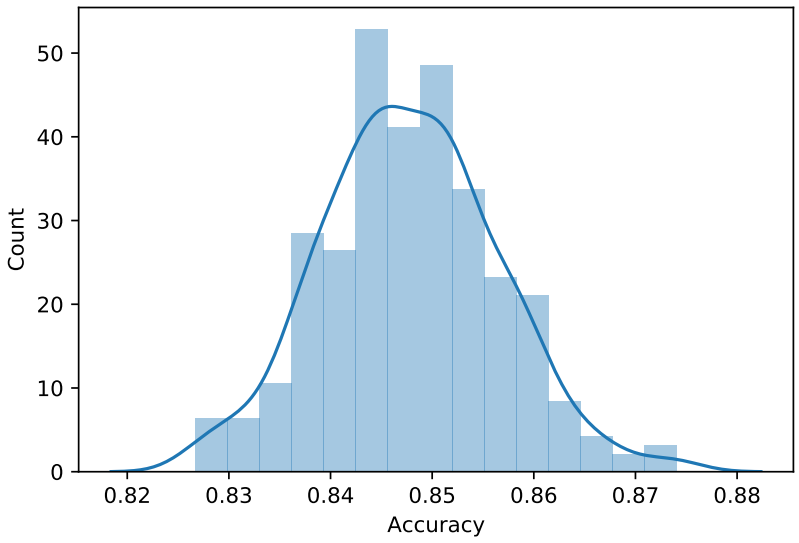

Fig. 4: Distribution of the accuracy in hold-3-out Prediction

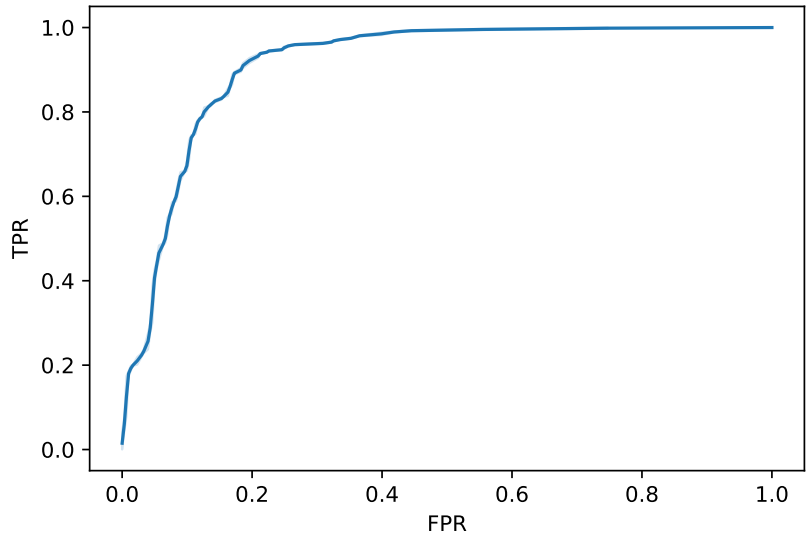

Fig. 5: AUC in hold-3-out Prediction

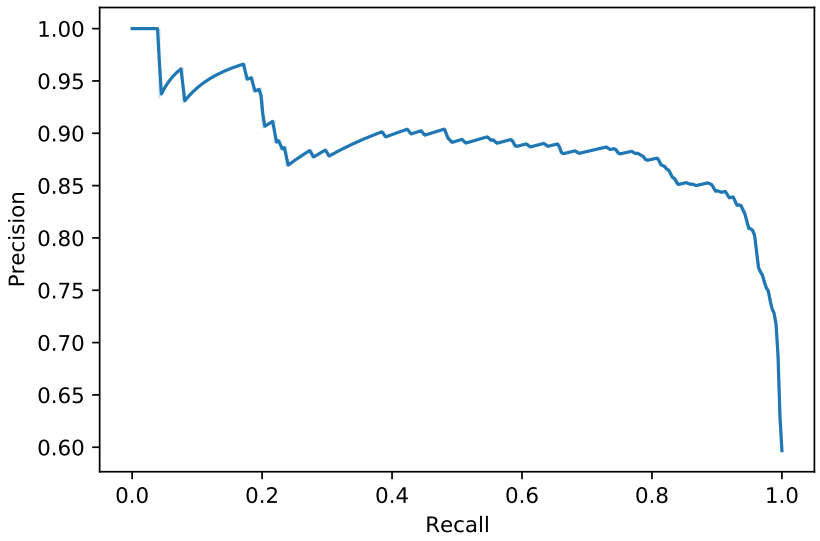

Fig. 6: ROC in hold-3-out Prediction

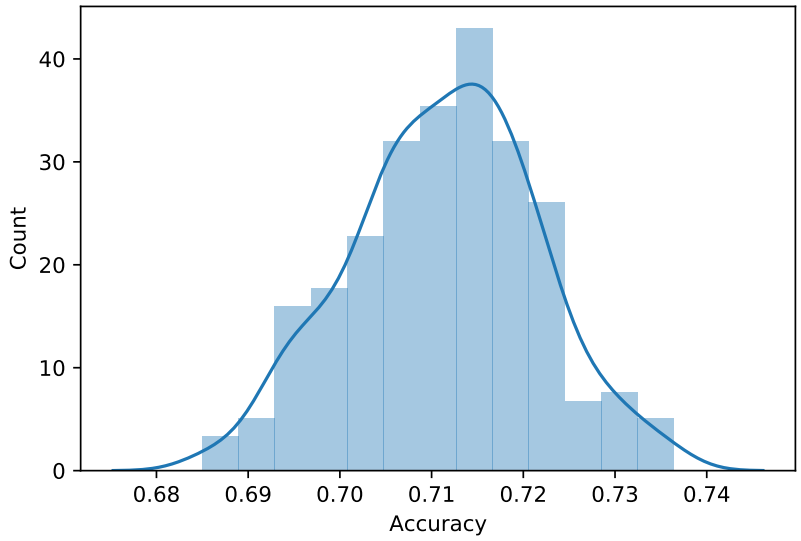

Fig. 7: Distribution of the accuracy in hold-9-out Prediction

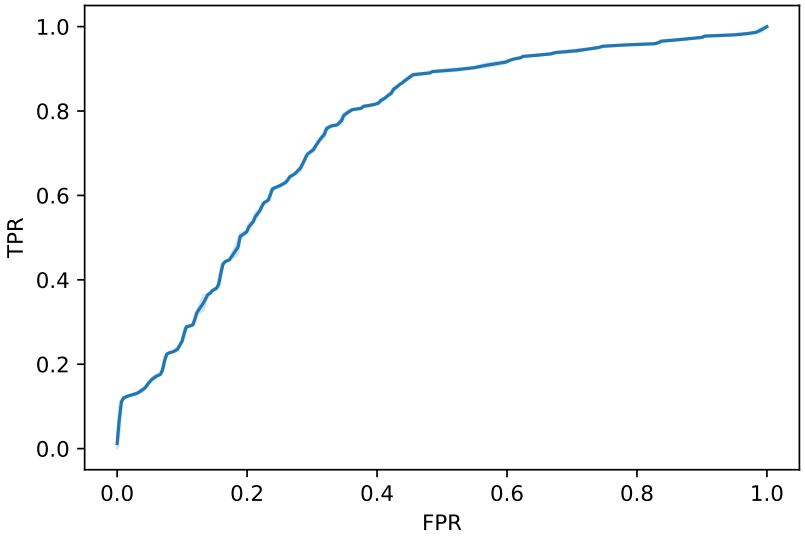

Fig. 8: AUC in hold-3-out Prediction

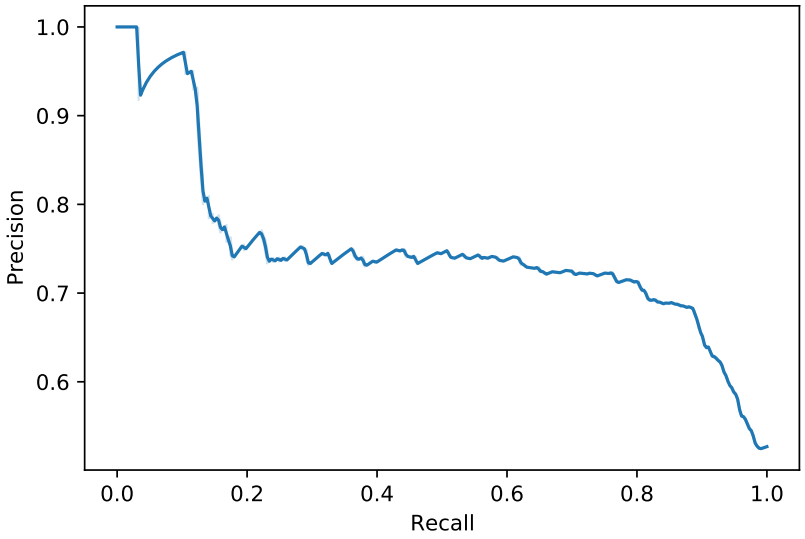

Fig. 9: ROC in hold-9-out Prediction

Model: "sequential"

Layer (type)	Output Shape	Param #
layer1 (Dense)	(None, 10)	110
layer2 (Dense)	(None, 10)	110
layer3 (Dense)	(None, 10)	110
layer4 (Dense)	(None, 1)	11

Total params: 341
Trainable params: 341
Non-trainable params: 0

Fig. 10: Neural Network architecture.

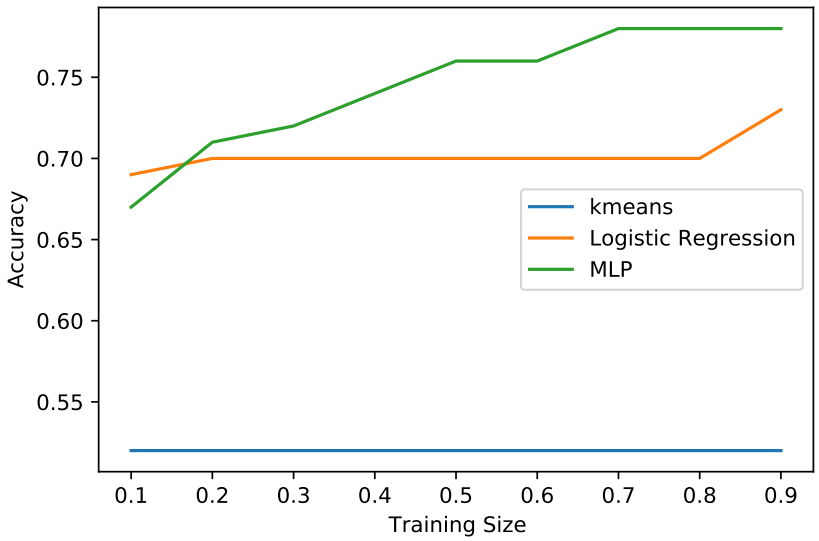

Fig. 11: Performance of different model based on the training size.