**Business Value:**

Demonstrating the possibility of uncertainty quantification modelling with VAE (Variational AutoEncoder) and LSTM embedding.This allows the model to have more flexibility in training to prevent the model to overfit while introduced some variations in the model through VAE.

This model will be able to predict different results as the latent space is sampled with different values. By running the prediction n times (say 100 times), we will be able to quantify our prediction risk, which is not available in point estimations, and will be able to use it to generate the prediction distribution and adjust the gamling odds accordingly.

It is also possible to scale the model to optimize integer outputs by changing the target variable to integers such as #corners, # of total scores, final game score etc. with probabilistic output and Monte Carlo simulation which is not covered in this demo but is discussed in the `suggested next step` section.

**Model Target**: Predict W/D/L for the games in the season (exclude first 5 games)

**Why the Model Strucutre**

Unlike normal models, everytime we sample from the Latent Space with a distribution so we will always get a different result.

We will be able to run through the prediction pipeline for say 30times to get the estimated prediction distribution.With the prediction distribution, we will be able to compute the mean and s.d. and can be further used in downstream tasks such as bet pricing.

**Train/Val/Test split:**

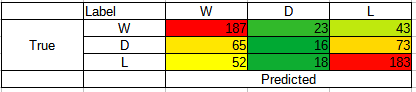
3:1:1 (2015-2017 for training, 2018 validation, 2018 testing)

Where 2014 data is used to get the statistics for 2015 standard\_scaler to prevent same year data is used to normalise the columns hence leading to use of future data.

**Results**

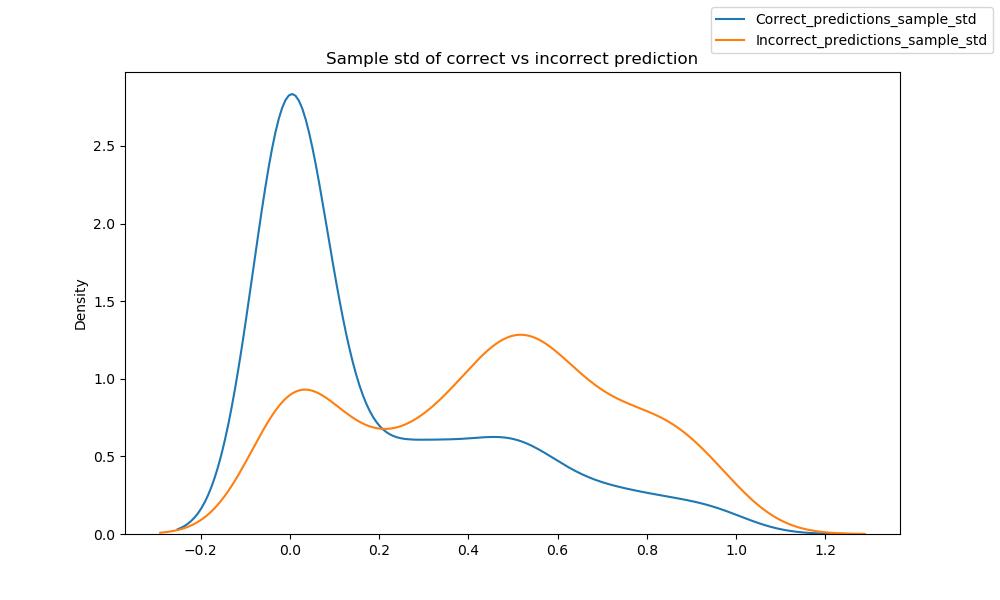
Win/Draw/Lose accuracy = 58.5%

Confusion Matrix:



(Larger numbers go with Red, smaller numbers go with Green)

**Key takeaway**:



The above figure demonstrated the key delivery of the model, uncertainty quantification.

The model has been used to predict 100 times with the samples drawn in VAE latent space each with different output. The above chart shows the standard derivations of the predicted label in the *test set with 2019 data* (smaller the better as it is distribution of std.). The blue line is the std distribution of all correctly predicted values and the orange is the incorrect ones.

**It is clear that the blue line distribution outperforms the orange one as it has a significantly lower std value.** Vice versa, we will be able to identify which prediction the model is more confident in the same way.

You may also check the histogram plots under `Results` folder in the zip file.

**Suggestion next steps:**

\*This section will focus on what we can do more with this model structure

1. Idea can scale to other outcome:

Without changing the input, the model can easily change into :

* "scored" and "conceded" with poisson distribution output, used for "total scored", "odd/even total scored", "First score team", "Half game result’’

By changing the input, the model similar structure can be applied to:

- Possession, # of balls passed, # of fouls for players, "# corners" etc.

1. Probabilistic prediction can also be applied (with say Tensorflow Probability) to allow a better "odds estimation" for different overall results

* Tensorflow released a product called tfp (Tensorflow Probability). With this model, we will be able to optimize the prediction by considering the target as a random variable.
* Result of this model will be backed by more "statistical support" and will be quantify

1. Other model training techniques can also be applied

* Other technique such as decay learning rate, dropout, layer\_normalization etc. can also added to the model to improve performance

**Concerns:**

\*This section will focus on the flaws on the model structure and give suggestions if more data/time is available

1. This model is focused on demonstrating the possibility of uncertainty quantification modelling but does not accommodate the performance of the first 5 games.

* Such predictions can be made using player level day of the past 5 games (with the stats provided in `Player\_Data` but in game level, even with the stats in the last 5 games in last season)

1. The model above does not leverage a lot of in depth data as it is not available in the provided data.

* A probabilistic model with VAE and distribution aliked output is suggested when more game level data is available.Similar approach can be used liked the lower part of the LSTM-VAE encoder model where first component can be player level stats (e.g. # passes, # shoot attempts, #assists, mins in penalty area etc.) and second level can be game level stats (possession, # fouls, # cornerns, #pass, etc.)

1. 1 Year data is completely dumped as it is used to compute standard\_scaler statistics

* The standard\_sclaer statistics can be replaced by computing with a rolling 5 games. On the other hand, with this methodology, the scaled metrics of the same team in the same year will not be scaled on the same level.