SN_DREAMS
Risk Factor Clustering using Autoencoders

Aadit Deshpande

Recap

- 13 Risk Factors for Diabetic Retinopathy (DR)
- 4 Categorical and 9 Continuous Columns
- Expert-Labelled Primary and Secondary Risk Factors
- Data points (Patients) with DR = 1 are sparse => Imbalanced Data
- 3 Approaches used before:
    1. K-means, Agglomerative Clustering
    2. KNN Classifier and ROC_AUC Score
    3. VAE and Weight Score

Modifications to previous code (1)

- Continuous Columns:
  - 4 risk factors
  - Range: >=0
  - Standard Scaler applied
- Categorical Columns:
  - 9 risk factors
  - Range: [0, 1]
  - No Standardization Applied
- Evaluation of Primary/Secondary prediction using a Confusion Matrix

Evaluation using Confusion Matrix

- Compare ‘Actual’ and ‘Predicted’ labels
- ‘Actual’ labels (Experts)
- Create Confusion Matrix
- Count TP, TN, FP, FN

Risk Factor	Actual	Predicted	Correctness
gen_cat	2	2	Correct
Hypertension_cat	1	2	Wrong
IHD_cat	2	2	Correct
insulin_treatment_cat	1	2	Wrong
fam_his_diab_cat	2	2	Correct
bp_sys	1	1	Correct
Physical_AS_cat	2	2	Correct
ann_family_in	2	2	Correct
HbA1c_grade_cat	1	1	Correct
agegrp_cat	2	1	Wrong
BMIcat	2	1	Wrong
DurYrs	1	1	Correct
FBS_Grp_cat	1	1	Correct

	Actual Primary	Actual Secondary
Predicted Primary	4	2
Predicted Secondary	2	5

Features in Correct clusters = 9
Features in Wrong clusters = 4

KNN Results

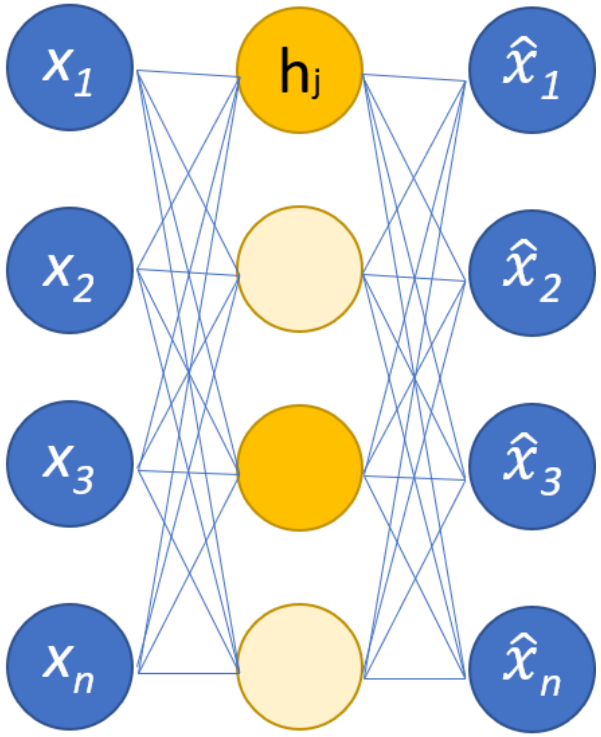
- KNN
  - SMOTE-ENN
  - 70:30 Train Test Split
  - K = 5
  - Metrics: ROC-AUC Score and Classification Accuracy
- ***8 Incorrectly Classified Features***
  - 4 False Positives
  - 4 False Negatives

```
Wrongly Classified Features:
    gen_cat
    IHD_cat
    insulin_treatment_cat
    bp_sys
    HbA1c_grade_cat
    agegrp_cat
    BMIcat
    FBS_Grp_cat
```

```
Confusion Matrix:
+---------------------+----------------+------------------+
|                     | Actual Primary | Actual Secondary |
+---------------------+----------------+------------------+
|  Predicted Primary  |       2        |        4         |
| Predicted Secondary |       4        |        3         |
+---------------------+----------------+------------------+
Features in Correct clusters =  5
Features in Wrong clusters =  8
```

Intro to Autoencoders

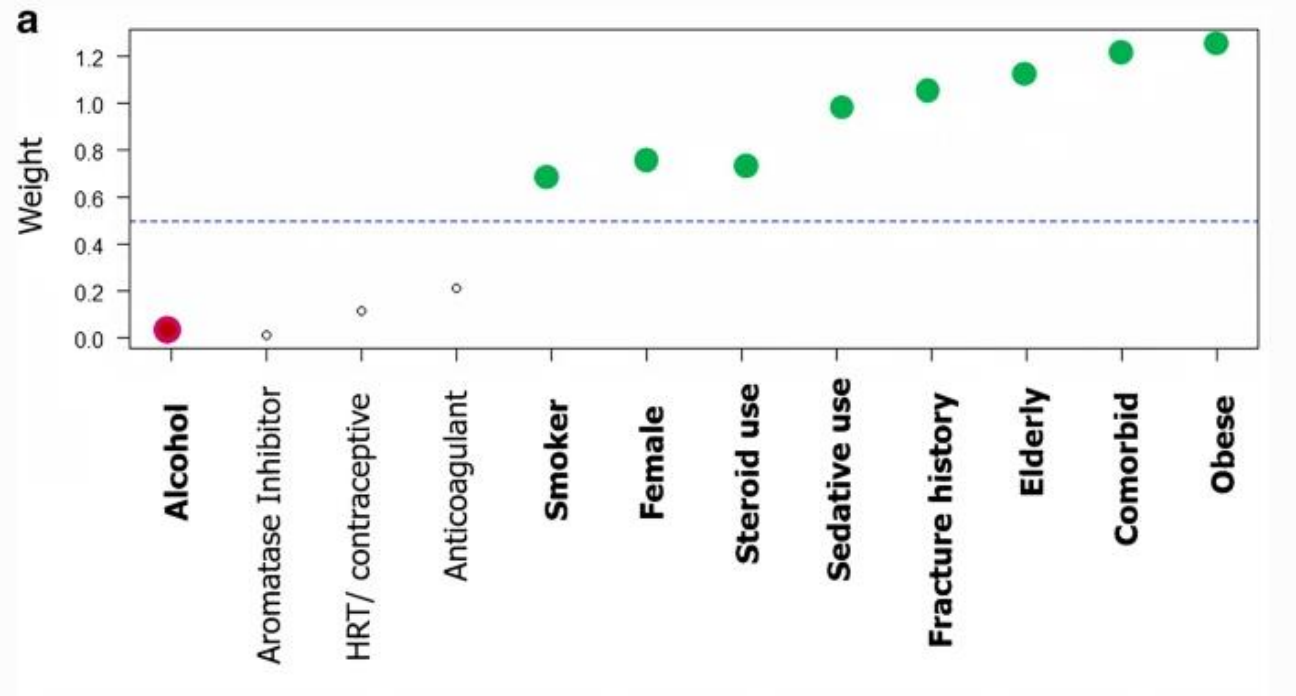
- Motivation: Using a Dense representation via Deep Learning may better be suited for Feature Selection (Primary/Secondary)
- **Autoencoder**
  - Artificial Neural Network
  - Learns representation (Encoding) from data
  - Validates, Refines the encoding
  - Used as generative models
  - Encoder: Maps [Input -> Code]
  - Decoder: Maps [Code -> Reconstructed Input]



Reference Paper

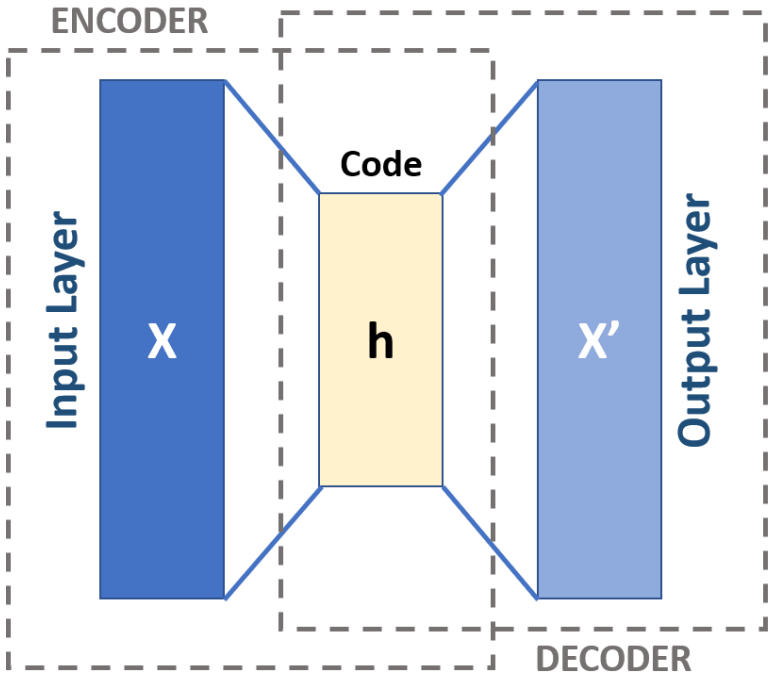
- Autoencoder Weights to select Features
- 8 risk factors Selected
- Same as those identified by clinical experts

Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research [Khalid et al. 2019]



Autoencoder (Attempt 3)

- 70:30 Train Test Split
- Standard Scaler (Only to continuous Columns)
- 2-layer Autoencoder
  - Densely Connected ***Code Layer*** (7 neurons)
  - Dense ***Output Layer*** (13 neurons)
  - 202 Parameters
- Training
  - Mean Absolute Error, Adam Optimizer, 15 epochs
- Metric: Weights Learnt by Code Layer to score features
- Threshold = Median (?)



Nature of Results

- Good Results most of the time (<=4 errors)
- However, lot of variation
- Threshold Selection
  A. Threshold
  B. Median
- Save Weights on a good result (<=2 results)
- Reload Weights (To eliminate training time)

Autoencoder Results (1)

Risk Factor	Actual	Predicted	Correctness
gen_cat	2	2	Correct
Hypertension_cat	1	2	Wrong
IHD_cat	2	2	Correct
insulin_treatment_cat	1	2	Wrong
fam_his_diab_cat	2	2	Correct
bp_sys	1	1	Correct
Physical_AS_cat	2	2	Correct
ann_family_in	2	2	Correct
HbA1c_grade_cat	1	1	Correct
agegrp_cat	2	1	Wrong
BMIcat	2	1	Wrong
DurYrs	1	1	Correct
FBS_Grp_cat	1	1	Correct

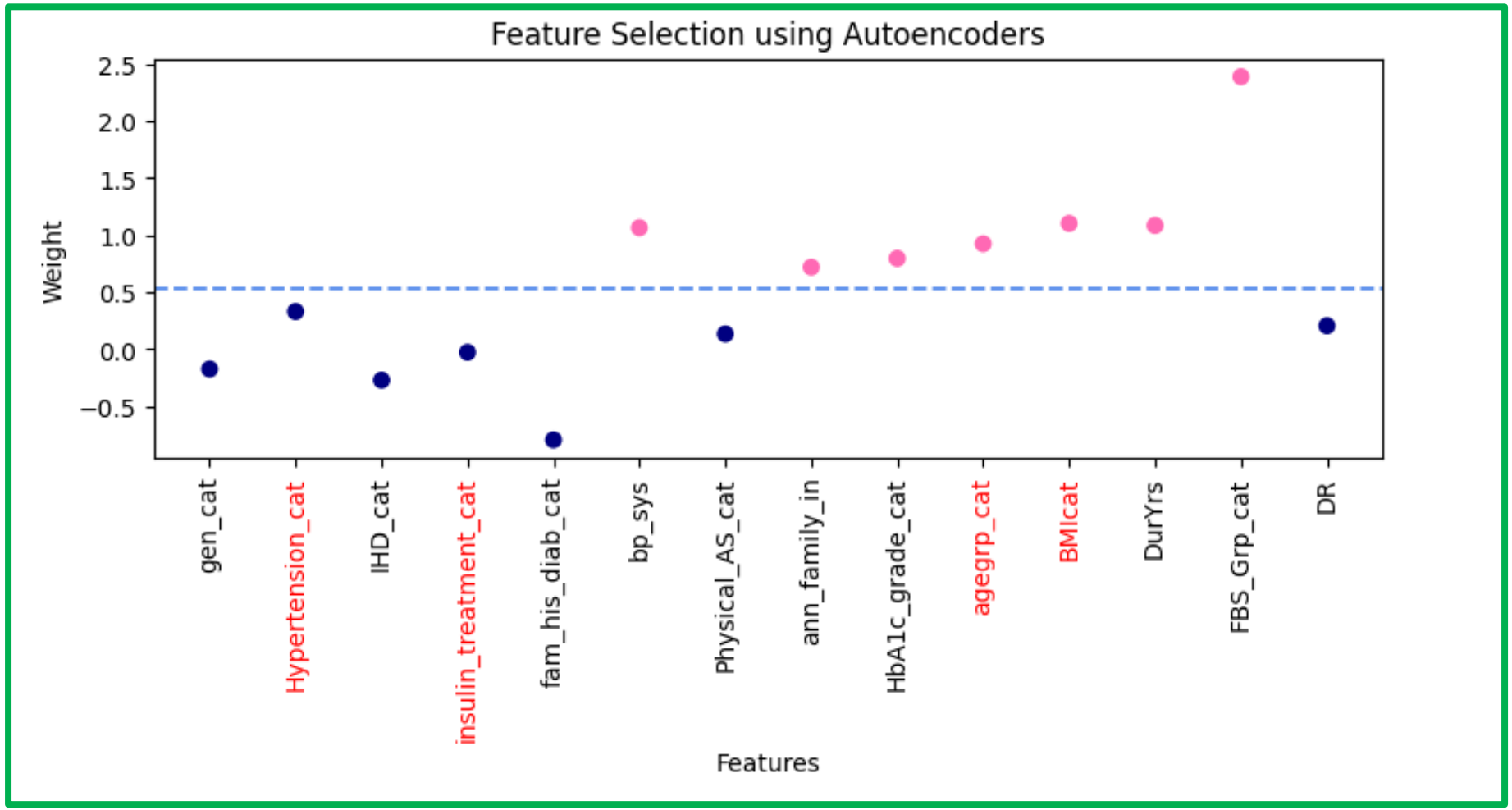
Wrongly Classified Features:
   Hypertension_cat
   insulin_treatment_cat
   agegrp_cat
   BMIcat

Confusion Matrix:

	Actual Primary	Actual Secondary
Predicted Primary	4	2
Predicted Secondary	2	5

Features in Correct clusters = 9
Features in Wrong clusters = 4

Autoencoder Results (2)



References

- https://www.tensorflow.org/tutorials/generative/autoencoder
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster
- https://towardsdatascience.com/dimensionality-reduction-pca-versus-autoencoders-338fcaf3297d
- Khalid, S., Prieto-Alhambra, D. Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research. *Curr Epidemiol Rep* **6,** 364–372 (2019).