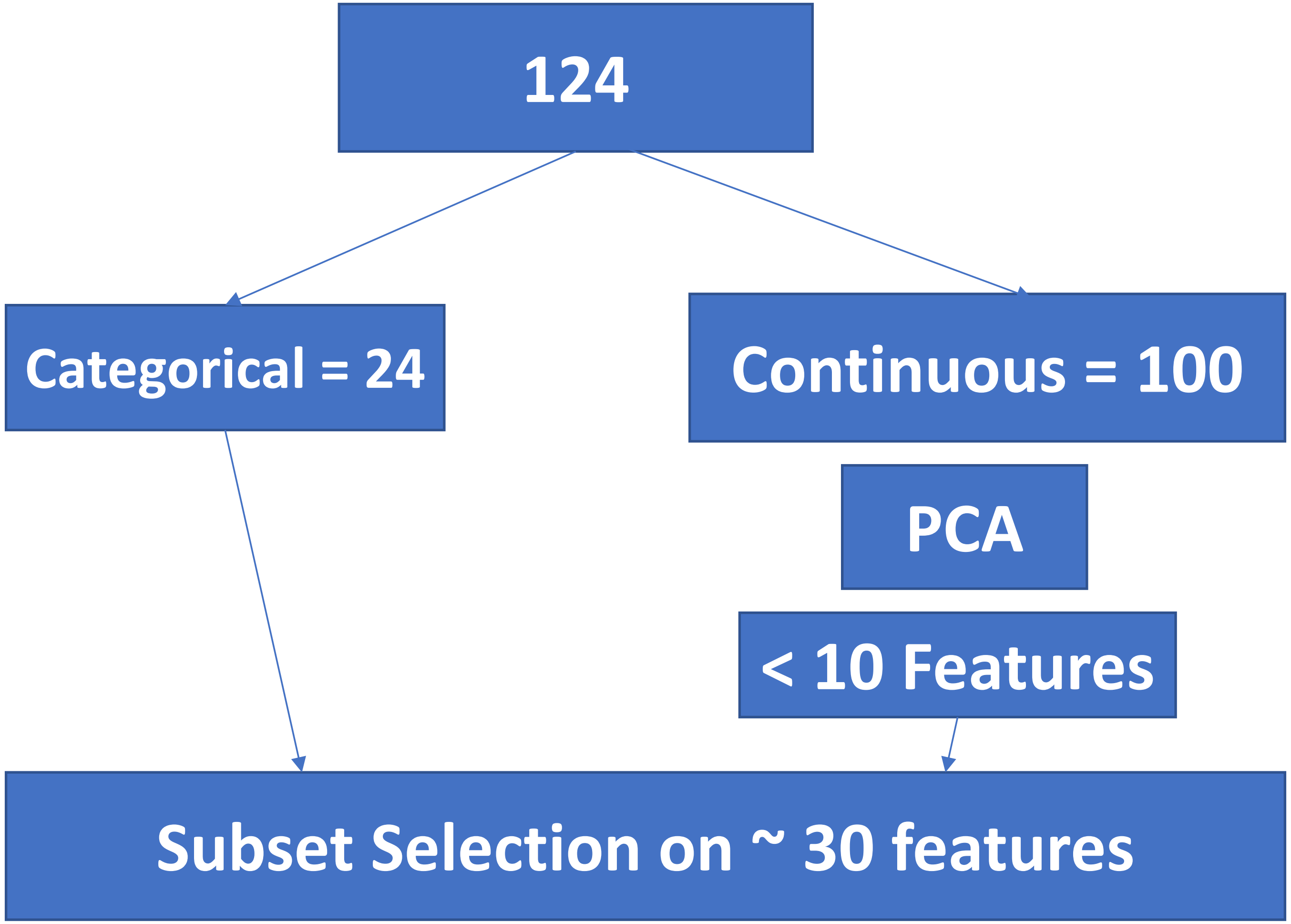
Project Group A13

Ankur Garg & Sanket Shahane

Plan of action



Estimated Time

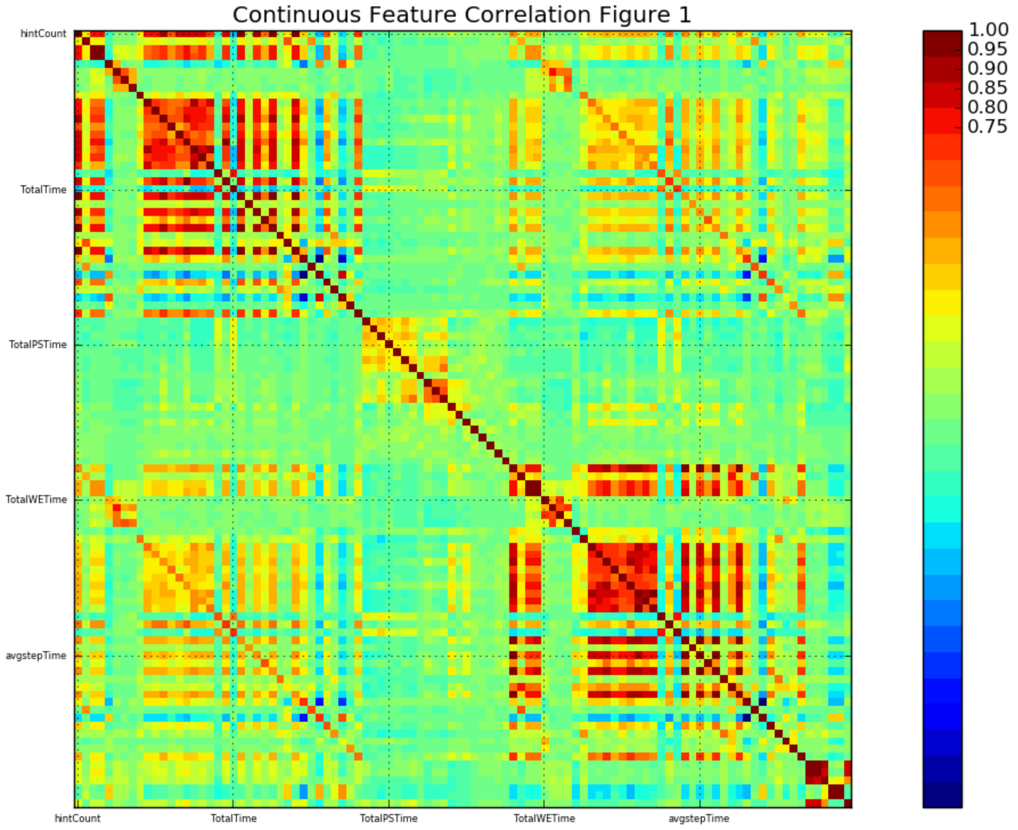
- 5-6 hours for computation

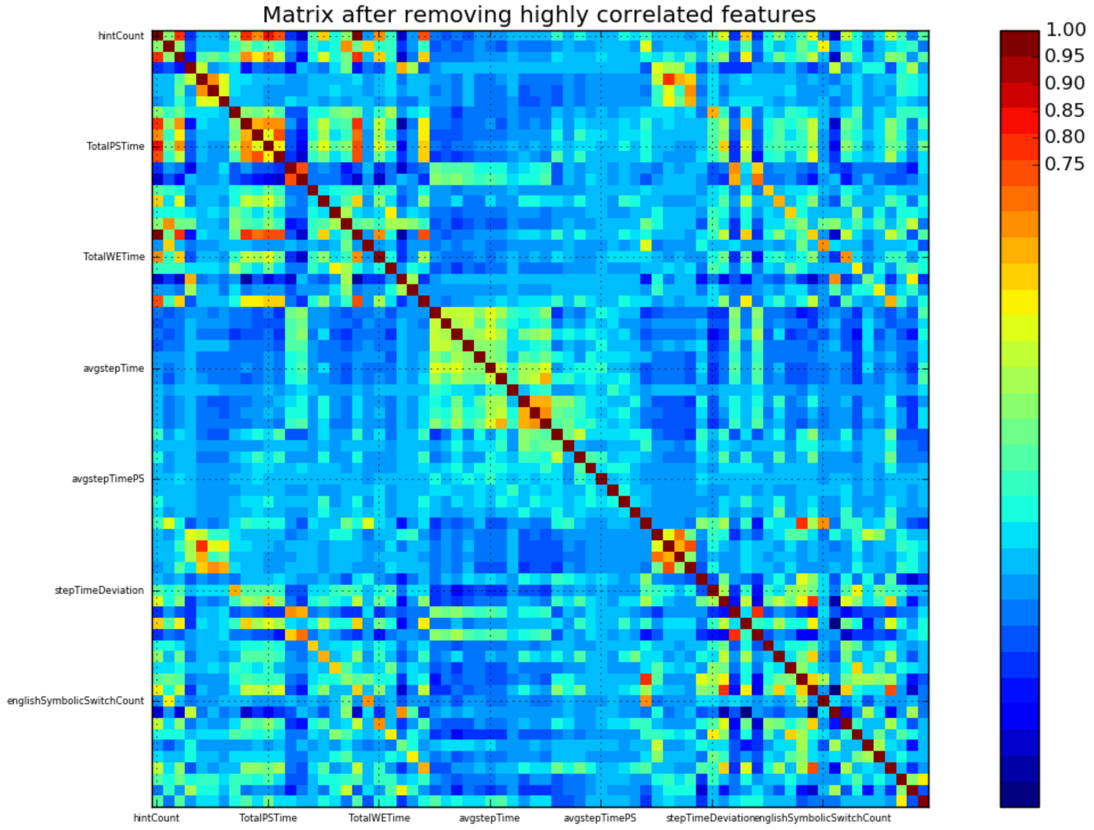
- Enjoy the weekend.

- However …

Handling Continuous Features

- Remove Correlated Variables before PCA (30 features removed)


Continuous Feature Correlation Figure 1


Matrix after removing highly correlated features

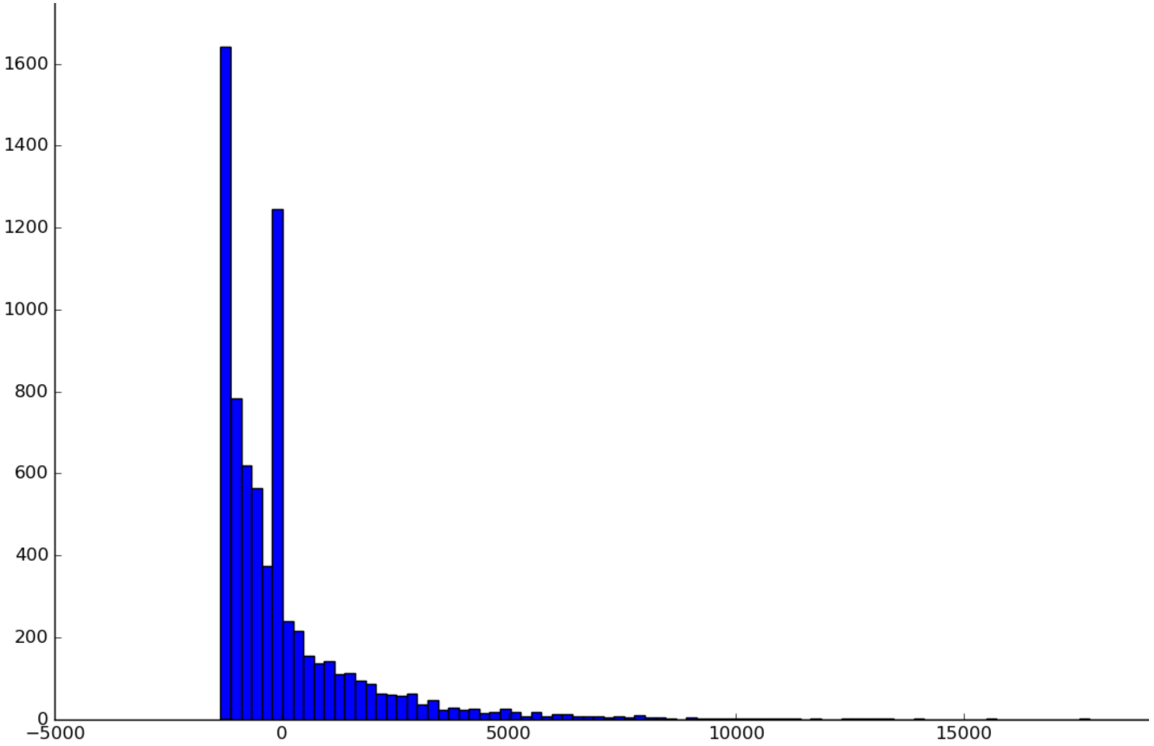
Principal Component Analysis

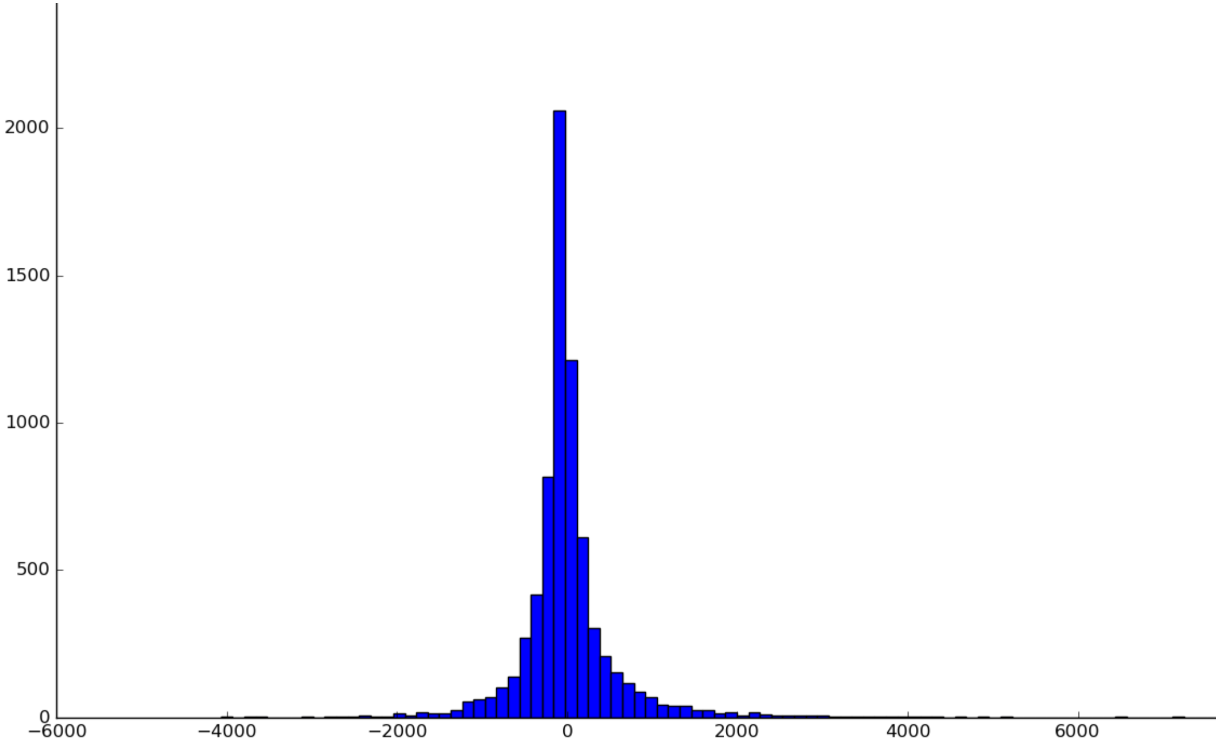
- Before removing correlated features, 3 components explained 97% variance

- After removal of correlated features, ended up choosing 6 principal components.

Discretization of Principal Components

1. Based on data distribution





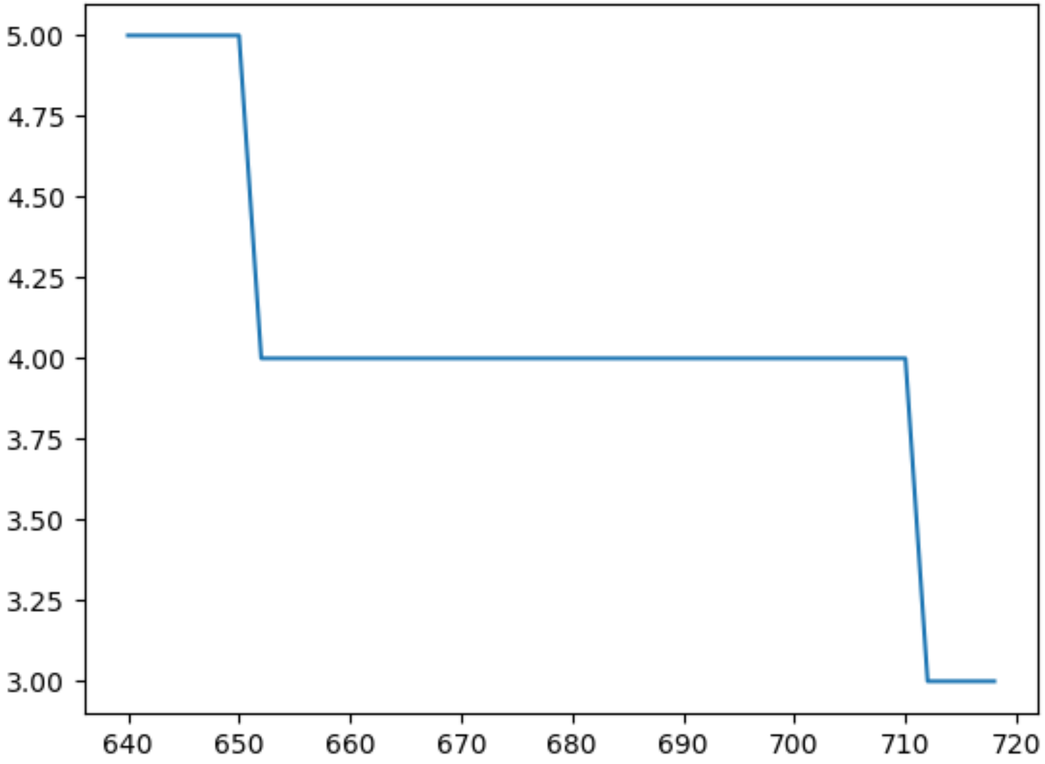
Discretization of Principal Components

2. Based on equal frequency

3. Equal width bins

Discretization of Principal Components

4. Clustering - using Mean Shift algorithm



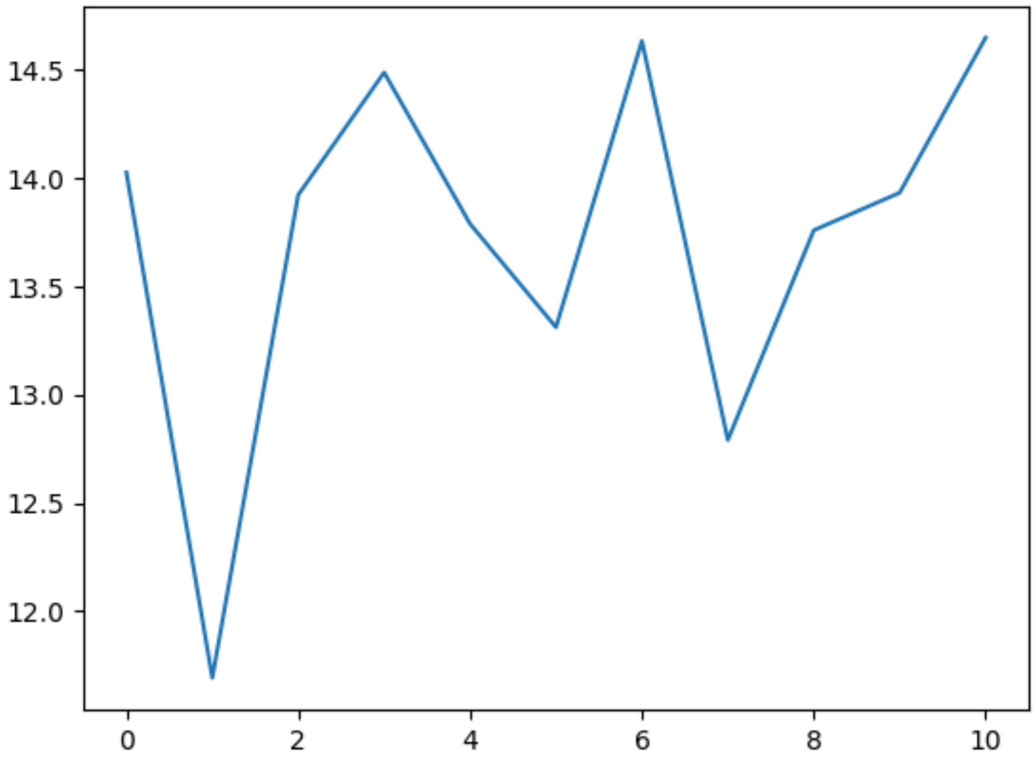
PCA Results

- Not so useful

- ECR values in the range of 0-15 using just discretized PCA components

PCA Results

- ECR for PCA components



Back to square one

- Forward stepwise subset selection

- Two good features – {Level, cumul_Interaction} - ECR value: 75.80

Improving ECR

- Greedy approach

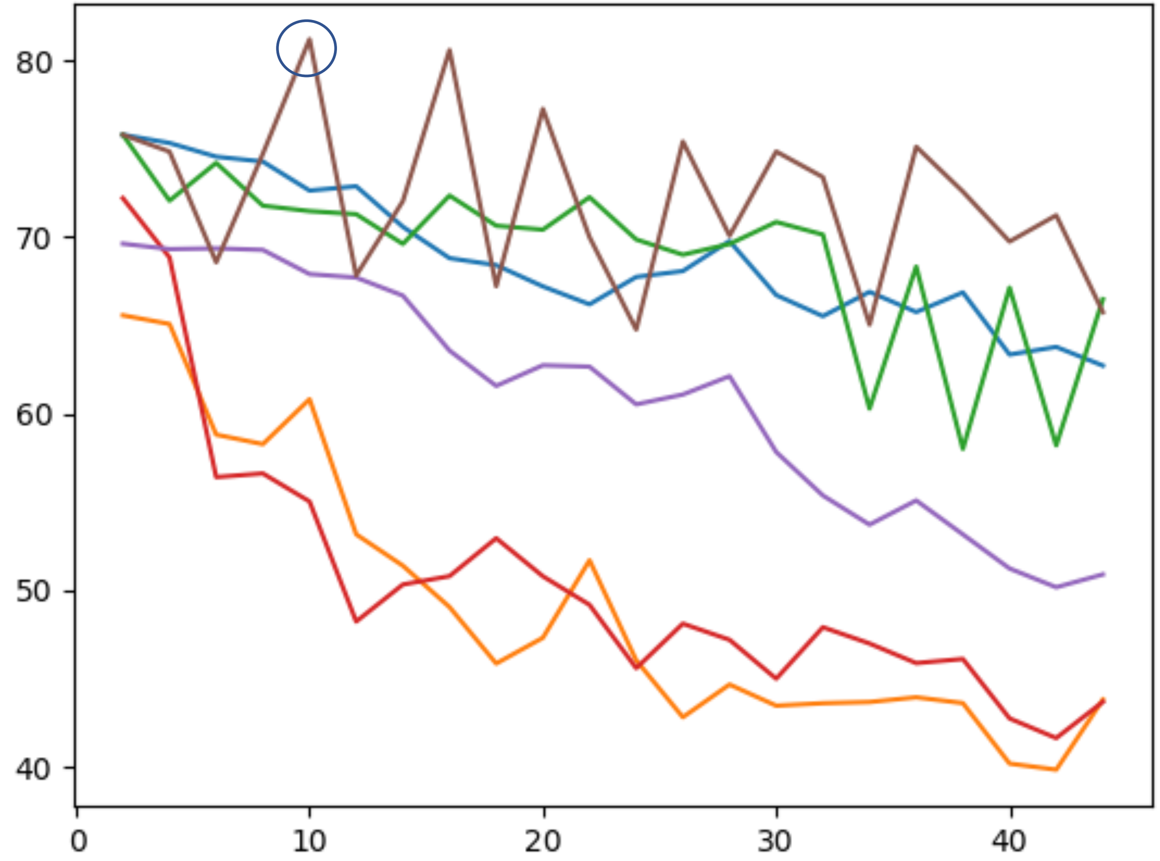
- Discretization of PCA and MFA components based on ECR value

PCA + Greedy discretization + Forward Subset

- PCA components along with features of forward subset selection
- Discretization of PCA components based on ECR value

PCA + Greedy discretization + Forward Subset

- Improvement in ECR – 81.26



Handling Categorical Features

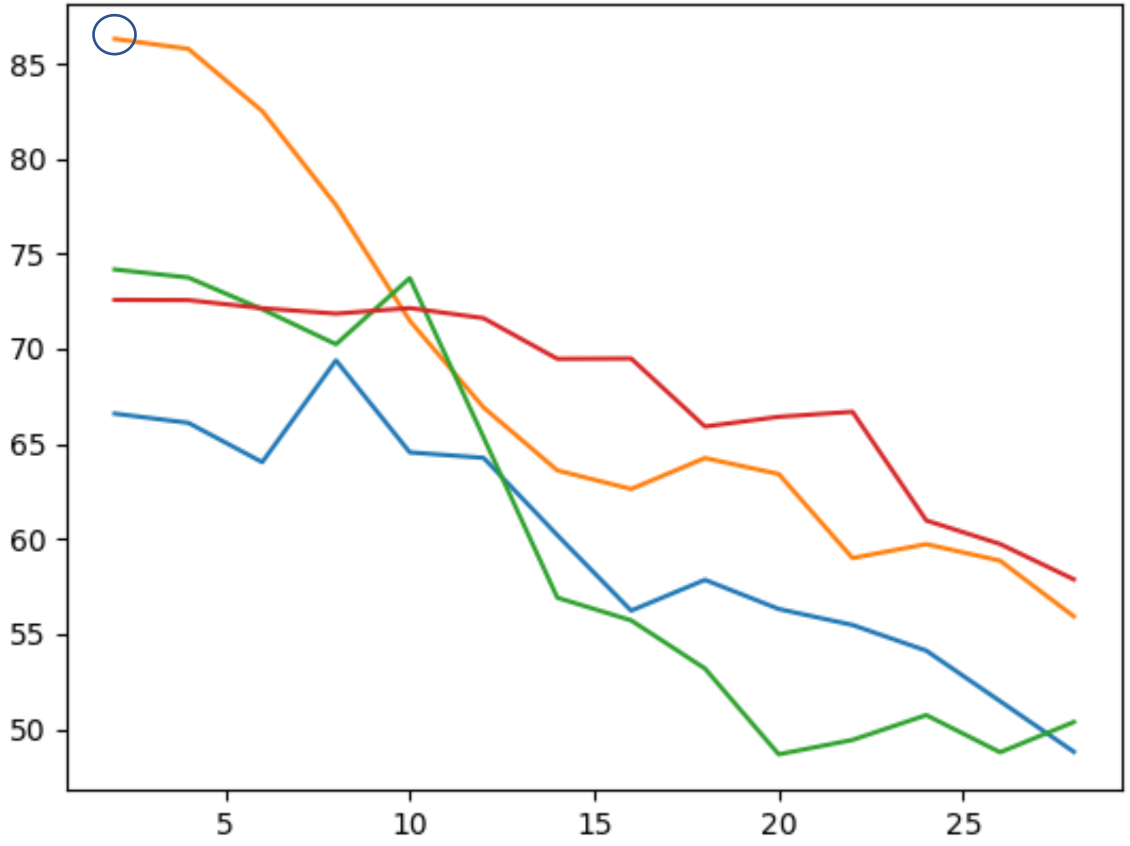
- Mixed Factor Analysis

- Extracted best 8 components

- Output of Mixed Factor Analysis is continuous
    - Discretized using previous methods

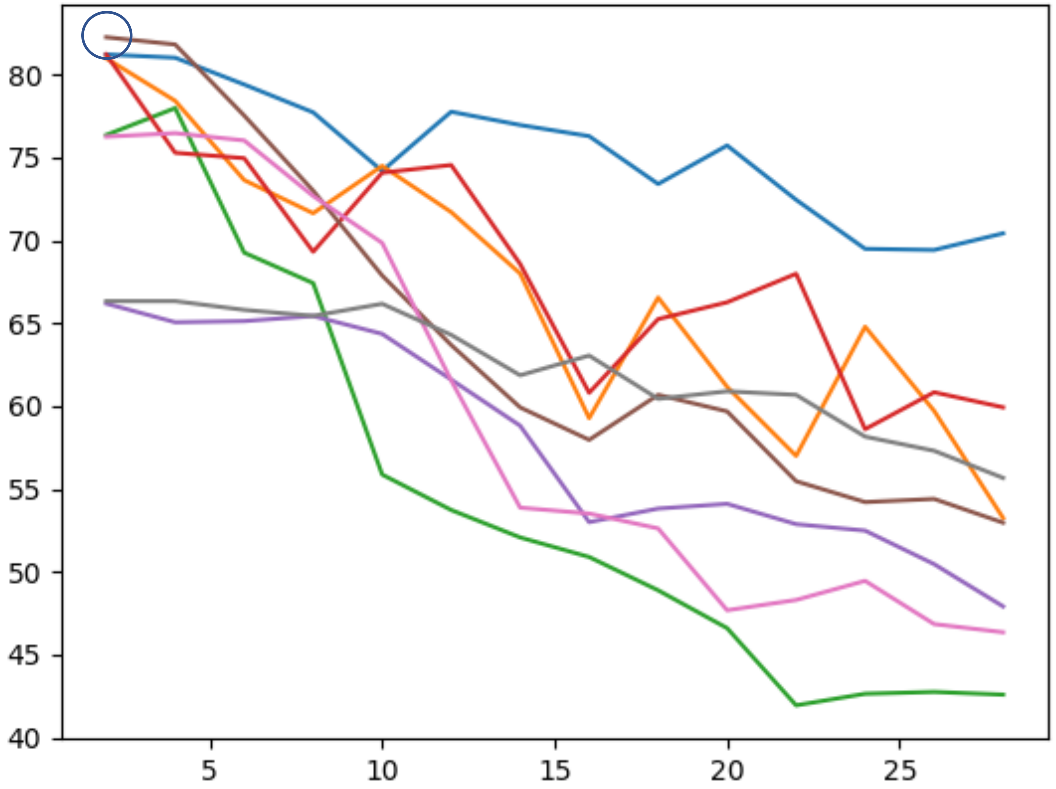
MFA + Greedy discretization + Forward Subset

- Improvement in ECR – 86.29



Combining results from PCA and MFA

- Using greedy approach to combine components with best results from both approache
- ECR Value - 82.5



Analyses

- Traditional feature extraction techniques don’t necessarily work in MDP
- Important to remove correlation
- Combining results from two good features doesn’t necessarily result in a good combination

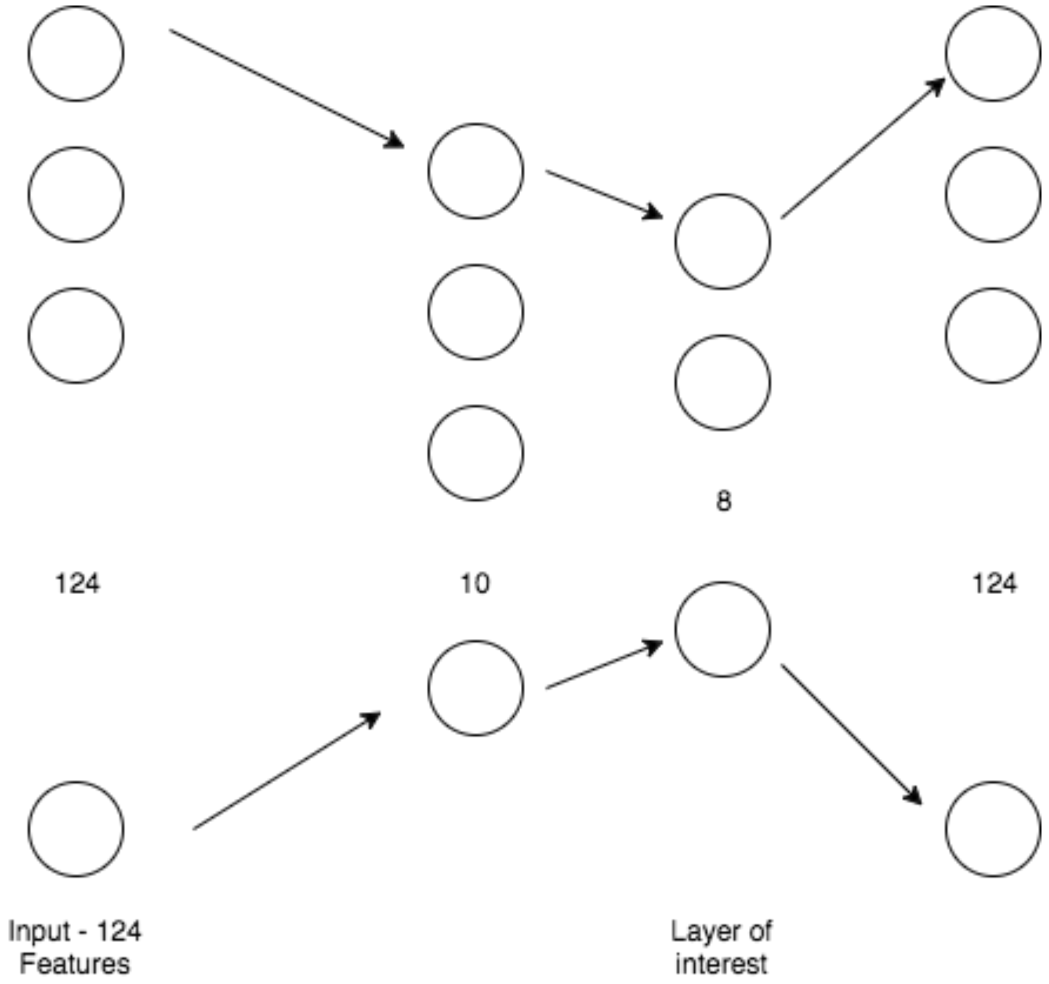
Another Approach

- Using Neural Network

- Method 1: Feature compression using neural network

- Method 2: Include MDP process while training the neural network

Neural Network - Structure

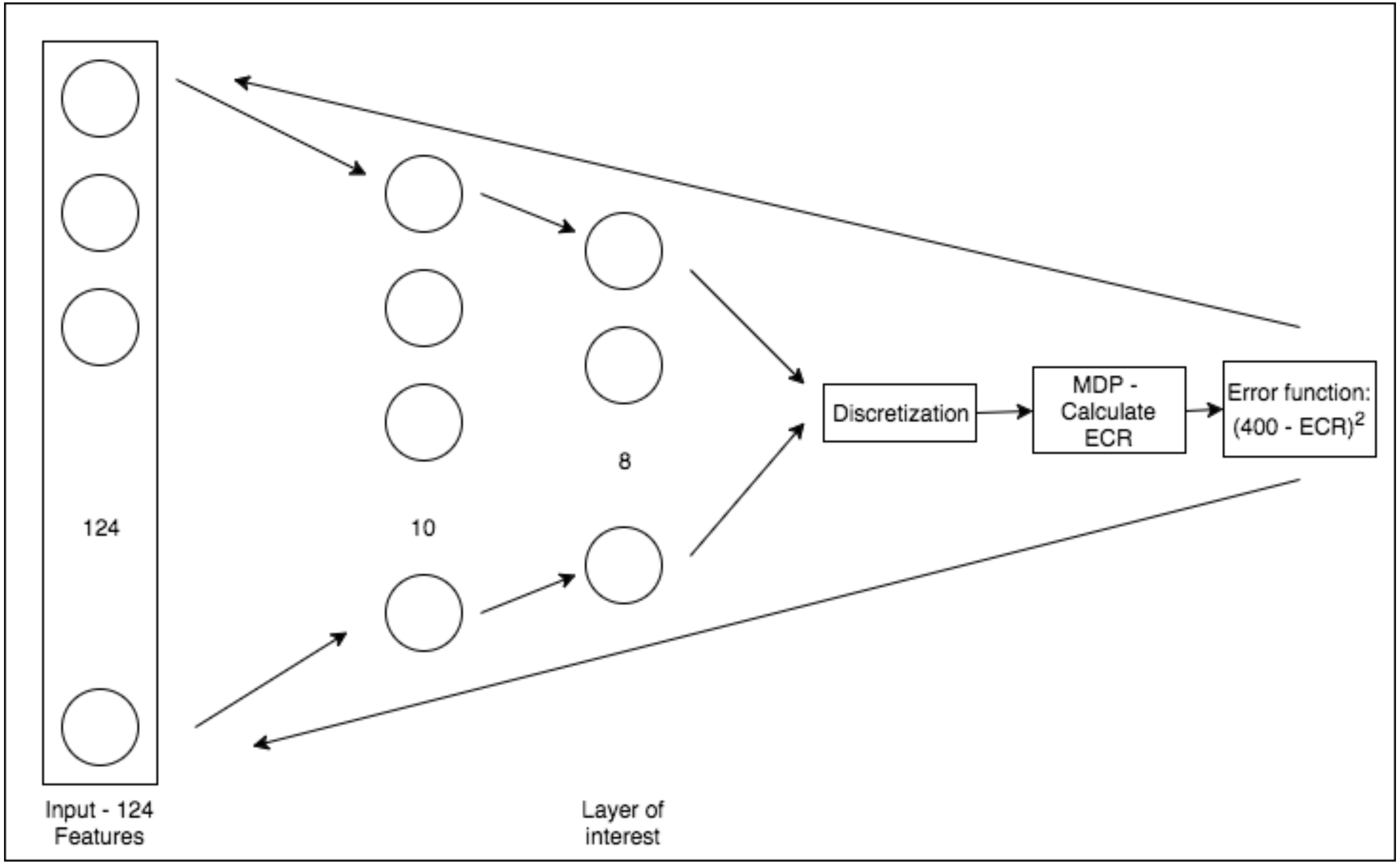


Neural Network - Results

- Extracted 8 features
- Discretized using previous methods

- ECR Value: 31.2

Method 2 – Using MDP for training NN



Final Results

- Max ECR till now – 86.29

- Total Rules: 3155

- WE Rules = 1073, PS Rules = 3155

Challenges and Outcomes

- Definition of good feature is rather difficult for MDP

- Explored different methods for feature selection and feature discretization

- Challenges in training and designing a neural network

- MDP_Function2.py is the bottleneck. Doesn’t execute in parallel.

Thank You

Group A13