Credit Card Default Prediction

Motivation

- Intrigued by the idea of influencing customer behavior positively to help them make better financial decisions.

- From personal experience, I have found it personally hard to manage credit card payments and I have ended up defaulting them multiple times.

- If had known at the start of a month that I would be potentially defaulting a credit card bill, I would have found ways to cut back on my spending.

- Without risk assessment on the repayment ability of customers, banks could potentially find themselves in crisis.

Problem Statement

1. To predict the likelihood of a credit card user defaulting using on a monthly payment based on a 6-month payment history and demographic features.

2. To discern patterns on the results derived from supervised learning and attempt to clusters the users into varying levels of risk.

Dataset

- The dataset contains payment information from October, 2005, from a reputed bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank.

- Among 30,000 observations, 5529 observations (22.12%) were records of credit card defaulters.
- There are two types of features in the dataset
  - Demographic Features - Age, Gender, Education, and Marital Status
  - Customer Financial Behavior Features - Credit Card Limit, payment history and bill statements from the last 6 months.

Evaluation Setup

- Data would be split into train and validation randomly.
    - Train - Validation set split - (25,000 - 5000)

- Key Performance Indicators - Supervised Learning (Order of Precedence)
    - F-1 Score
    - AUC
    - False Positive Rate
    - Accuracy

- Key Performance Indicators - Unsupervised Learning (Order of Precedence)
    - Elbow curves for number of distinct clusters
    - Silhouette scores

Results Supervised Learning Model

KPI	Logistic Regression	Random Forest	Gradient Boosting	Neural Network
F1 Score	0.37	0.47	0.45	0.51
AUC	0.61	0.65	0.65	0.68
FPR (%)	2.41	3.70	3.88	5.91
Accuracy (%)	82.10	83.26	82.76	82.90
Precision (%)	73.09	71.54	69.52	65.17
Recall (%)	24.39	34.69	32.99	41.21

Neural Network (Adam optimizer and 5 hidden layers with dropout layers) outperforms all the traditional machine learning model after substantial tuning.

Tuning Neural Network

Switch to Jupyter Notebook - Link

Unsupervised Learning - Procedure

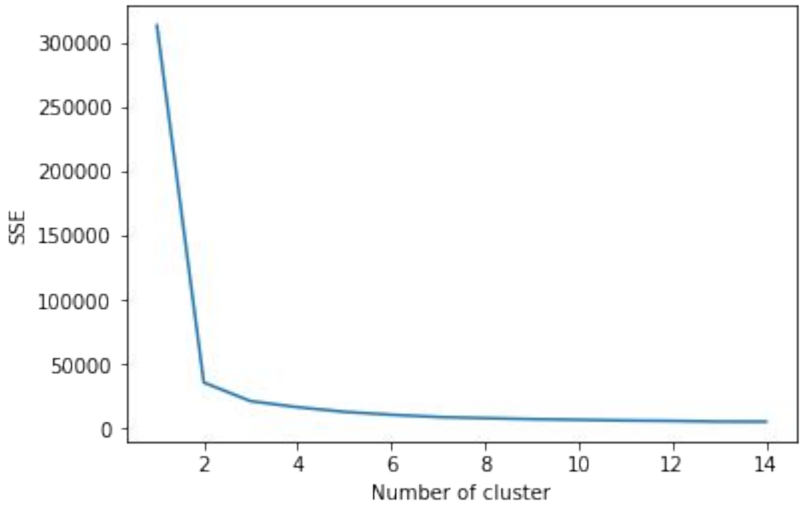
1. Use predictions made on the test set by the best model (neural network with adam optimizer and 6 hidden layers).

2. Decompose the X- feature space (84 dimensions) into 2 principal components for easier evaluation as some features are one-hot encoded.

3. Use elbow curve to figure out the appropriate number of clusters for the dataset.

4. Use silhouette score to ratify the significance of clusters.

Unsupervised Learning

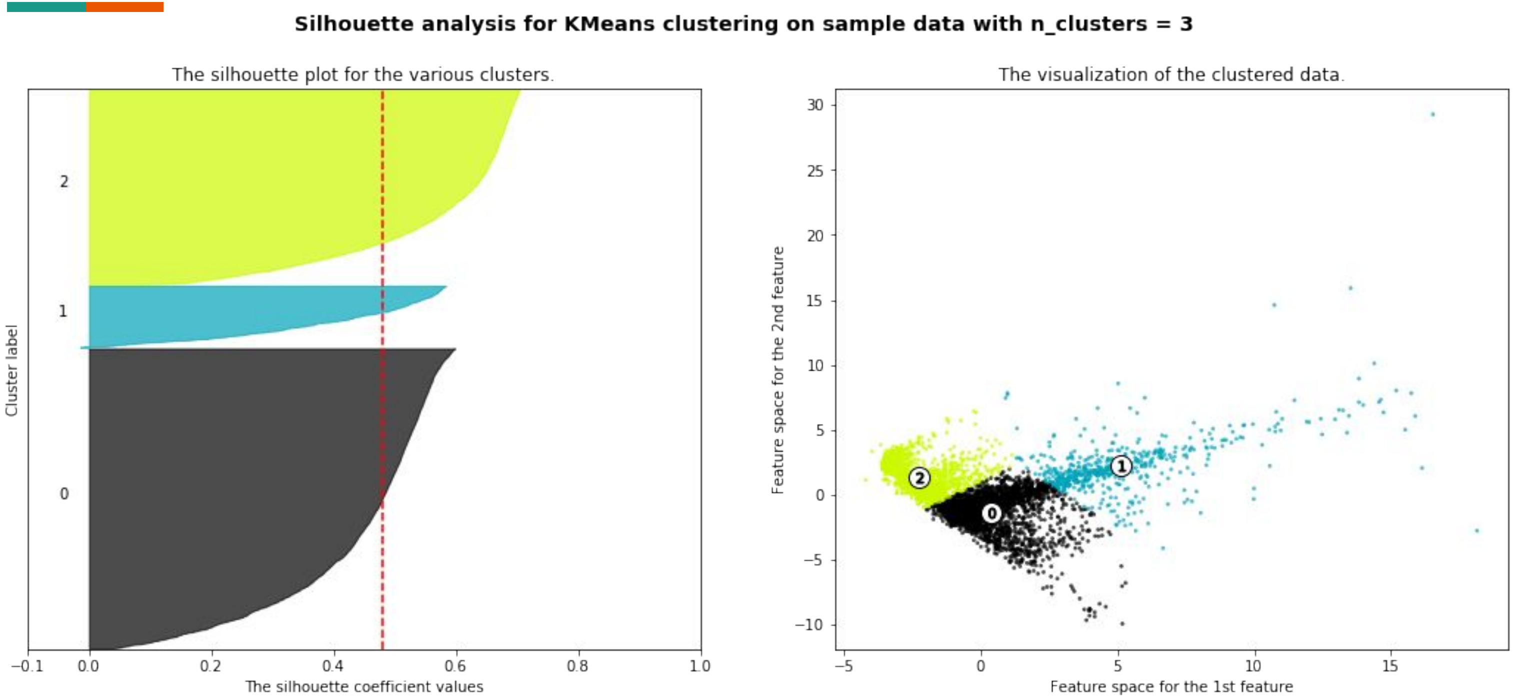


Elbow Curve

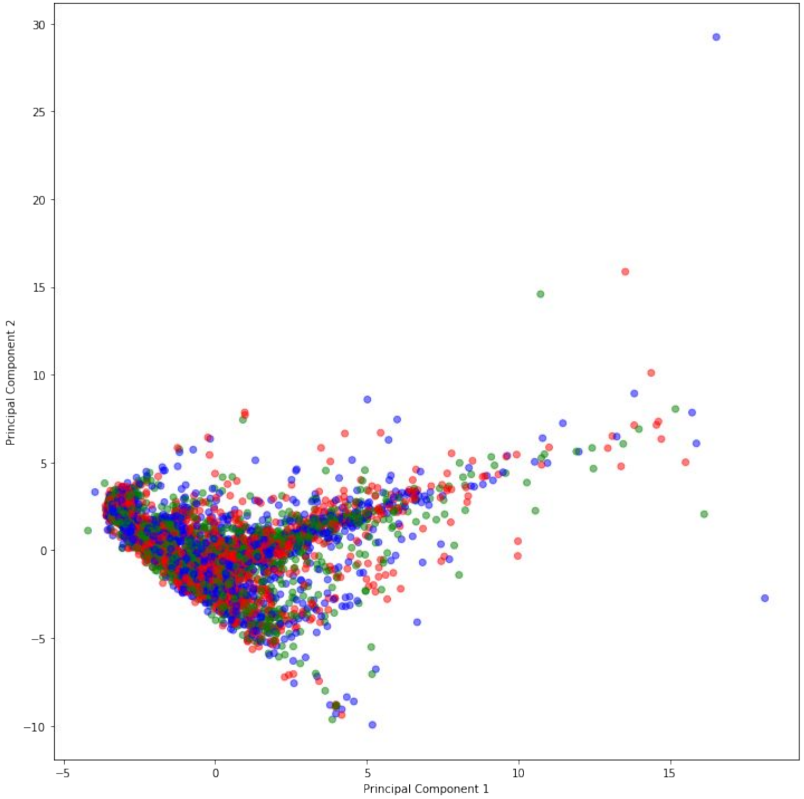
Number of Clusters	Silhouette Score
3	0.480457047
4	0.468868967
5	0.472492888
6	0.448608291
7	0.444949954

Silhouette Score

Unsupervised Learning Results



Unsupervised Learning - Percentile Based Clusters



Conclusions

- Neural Nets performed the best with a F1 score of 0.51 and a AUC of 0.67

- Based on the predictions, it was possible to segment the user group into 3 distinct groups although it's hard to interpret the clusters.

- With more features like FICO scores or household income, the accuracy of the model can definitely improve and the clusters will become more interpretable. Payment date related data, financial prudence