CS6890: Fraud Analytics - Assignment 3
Identifying outliers in the data by using Variational Autoencoders

Akshay Santoshi - CS21BTECH11012
Nitya Bhamidipaty - CS21BTECH11041

April 2025

1 Introduction

In this assignment, we applied a Variational Autoencoder (VAE) to a dataset (data.csv) followed by clustering analysis to identify outliers. The dataset consists of 10 features across 1199 samples. The goal was to generate bad points by performing clustering on the latent representations to detect outliers using k-means on non-linear representations.

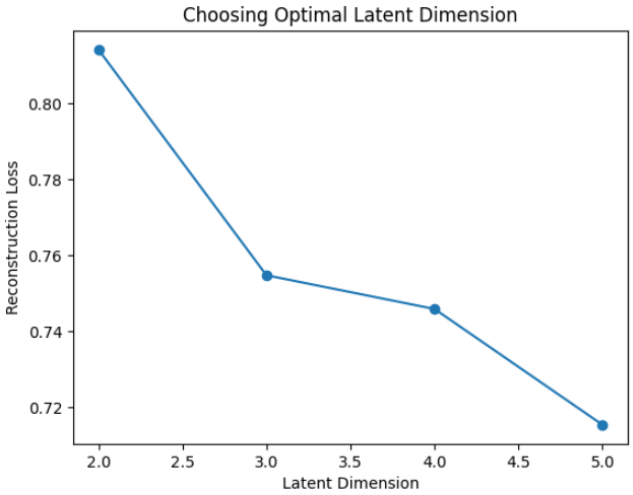
2 Approach

- First, we did data preprocessing, where we normalized the data. This is important for VAE training.
- Next, we defined a VAE model which consists of the encoder and decoder. For encoder, we used three linear layers, leakyRELU activations and batch normalization. We used similar architecture for decoder.
- For latent dimension selection, we tested multiple latent dimensions (2, 3, 4, 5) to determine the optimal one. The choice of latent dimension was taken to be 5 after observing the plot of latent dimension vs reconstruction loss. We used this value to train the VAE.
- Synthetic data was generated from sampling. This was evaluated by comparing feature distributions (KDE plots) and correlation heatmaps to assess similarity.
- Next, we applied K-means clustering on the latent representations obtained from encoder. We used elbow method to determine the optimal value of k (number of clusters).

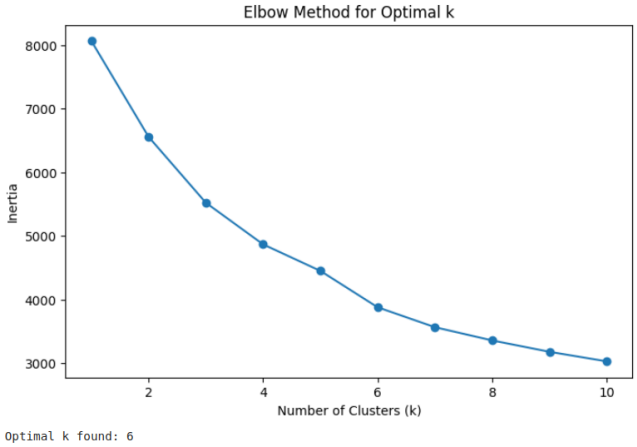
- For outliers detection, we considered small clusters which are clusters with fewer than 50 points as outliers. Additionally, the boundary points of big clusters, which are points beyond 95 percent of distance from the cluster center were identified as outliers.

3 Results

- We chose latent dimension for VAE as 5, after testing for loss in dimensons: [2, 3, 4, 5].



- Using elbow method to find the optimal number of clusters for kmeans, we got k = 6.



- In big clusters, we got the boundary points as shown in the table below:

=== Boundary Points from big clusters ===

	cov1	cov2	cov3	cov4	cov5	cov6	cov7	sal_pur_rat	igst_itc_tot_itc_rat	lib_igst_itc_rat
5	0.595378	-0.531958	0.679654	-0.126799	0.455487	0.432046	0.988092	-0.029813	0.768742	-0.054167
36	0.876536	0.974149	0.895181	0.750179	0.441621	-0.277212	0.801766	-0.031696	0.257575	-0.053998
45	0.188614	0.031166	0.832167	0.342503	-0.153480	0.824574	0.152444	-0.013936	-0.296467	-0.052397
53	0.999971	0.801451	0.450078	0.296673	-0.671271	0.221535	0.935767	-0.032791	0.313554	-0.054196
60	0.995746	0.943895	-0.294339	-0.409825	-0.546041	0.567263	0.850460	-0.032687	0.143176	-0.054144
64	0.996391	0.960214	-0.661415	-0.755766	0.494494	0.980598	-0.061126	-0.032326	-0.951539	-0.050436
65	0.989749	0.939645	0.973857	0.944348	0.450608	-0.020708	0.958587	0.022532	0.704163	-0.053252
99	0.950376	0.871910	0.268299	0.261641	0.171107	-0.574856	0.411161	-0.032145	0.302408	-0.054163
102	0.999988	0.998002	-0.263647	-0.279467	0.000000	0.490375	0.356098	-0.033177	-1.066401	7.486296
106	-0.133945	0.283608	0.970914	-0.171278	0.000000	0.976835	0.905625	-0.030864	-0.375026	-0.053857
124	0.946262	0.741639	-0.122708	-0.149332	-0.450775	0.507202	0.880061	-0.031580	0.924464	-0.054280
161	0.999999	0.973877	0.270032	0.223560	0.000000	0.066139	-0.225197	-0.030177	-0.962634	-0.047378
201	0.033564	-0.146953	-0.142238	-0.215256	-0.133407	0.653039	0.547611	-0.029229	-0.849317	-0.052171
202	1.000000	0.903017	-0.665221	-0.674896	0.000000	0.947314	0.591421	-0.032109	-1.066436	33.188277
213	0.968639	0.987501	0.821807	0.740876	0.000000	-0.164718	0.517192	-0.031127	-0.692171	-0.052769
249	1.000000	0.856922	0.823258	0.497245	0.000000	0.591521	0.021841	-0.029940	-1.066299	4.959007
251	0.999778	0.938575	0.407715	0.469158	-0.719622	-0.682734	0.989535	-0.033599	1.397712	-0.054371
261	0.985373	0.981808	-0.549108	-0.458748	0.522045	0.719086	-0.132245	-0.032735	-0.497513	-0.053758
292	0.999974	0.994909	0.615741	0.612976	0.000000	-0.095598	0.055525	-0.032150	-1.056888	-0.001766
299	0.997345	0.996813	0.270609	0.200251	-0.307178	0.873153	0.800368	-0.031731	-0.293160	-0.053854
308	-0.166308	0.537869	0.177641	0.320768	-0.275142	0.393301	0.936062	0.011202	1.120954	-0.054245
320	0.992803	0.610588	0.467223	0.053855	-0.063026	0.879758	0.941892	-0.032067	1.372217	-0.054303
378	1.000000	0.985049	-0.818128	-0.839158	0.000000	0.943767	0.589556	-0.032756	-0.969197	-0.050059
409	0.995813	0.827854	0.159099	0.458682	0.296877	0.987769	0.995211	-0.032056	1.975558	-0.054350
416	1.000000	0.937775	0.518851	0.498893	0.000000	-0.480075	0.078438	-0.031209	-1.064084	0.205998
453	0.812004	0.197821	-0.324095	-0.327846	0.723411	0.326996	-0.348182	-0.031032	-0.612744	-0.053296
462	-0.312219	0.993686	0.546496	-0.083316	0.000000	0.317602	0.633745	-0.009160	-0.790955	-0.052700
479	1.000000	1.000000	0.103333	0.103335	0.000000	-0.202757	0.773134	-0.032024	-0.269869	-0.053854
484	0.897166	0.973910	0.124157	0.024682	0.000000	-0.253106	0.970804	-0.030682	1.419545	-0.054194
492	0.990779	0.999023	-0.743153	-0.792628	-0.401923	0.848454	0.894657	-0.033177	0.460462	-0.054247
495	0.937190	0.998131	-0.289857	-0.579922	-0.535282	-0.099784	0.999547	-0.033258	2.126933	-0.054385
504	1.000000	0.999613	0.000000	0.000000	0.000000	-0.196621	0.998893	-0.033182	2.136006	-0.054387
544	0.993797	0.626927	0.533724	0.078995	-0.421030	0.968341	0.054833	-0.032082	-1.046099	-0.028653
559	0.999995	0.999789	0.820507	0.819809	0.000000	0.477295	-0.641931	-0.032637	-0.682291	-0.053340
564	0.999991	0.985618	0.521109	0.517412	0.000000	-0.264617	0.331915	-0.032833	0.632495	-0.054257
573	0.981851	0.996076	0.861156	0.851760	0.000000	0.021453	0.258469	-0.028863	-0.959804	-0.046591
579	0.999665	0.999687	0.742898	0.736978	0.000000	-0.199314	0.293745	-0.032287	-0.136891	-0.053980
590	0.935332	0.994236	0.554837	0.554199	0.000000	-0.175750	0.487344	-0.032483	-0.991349	-0.048398
591	0.245619	0.999049	0.300700	-0.059870	0.000000	0.793884	0.786821	34.367195	0.391459	-0.054193
603	0.999075	0.998665	0.221312	0.216898	-0.709599	0.906405	0.960929	-0.032578	1.303369	-0.054316
611	0.893083	-0.280000	0.622090	0.222152	0.563242	0.908451	0.999338	-0.032304	2.170322	-0.054357
668	0.999562	0.942223	-0.488435	-0.645740	0.546253	0.978647	-0.072301	-0.032506	-1.063466	0.099426
719	0.995573	1.000000	-0.499830	-0.469569	0.000000	0.817349	-0.859529	-0.033682	-1.061556	0.033165
727	0.999814	0.999847	-0.142689	-0.150315	-0.585397	0.786954	0.136990	-0.032751	0.239111	-0.054174
742	1.000000	0.758236	0.520414	0.052706	-0.473722	0.947093	0.709609	-0.032642	-0.647499	-0.053441
758	1.000000	1.000000	-0.119516	-0.119513	0.000000	-0.334804	0.855163	-0.032522	-0.620944	-0.053473
834	0.999861	0.998006	-0.625201	-0.644975	0.000000	-0.626728	0.143720	-0.032908	0.702145	-0.054270
852	0.126200	0.258904	0.146875	-0.178838	0.000000	-0.213063	0.985947	-0.035313	0.914433	-0.054336
859	0.988302	0.899342	0.507492	0.467459	0.000000	-0.511736	0.940332	-0.033288	0.163180	-0.054198
866	0.999428	0.822074	0.400275	0.325327	0.000000	-0.248904	-0.230404	-0.031192	-1.025085	-0.039506
874	0.996589	0.999892	0.902333	0.880036	0.000000	0.871109	-0.520048	-0.031821	-1.064662	0.270302
935	0.973113	0.904080	-0.281079	-0.306501	0.000000	-0.201455	0.972672	-0.034253	0.161957	-0.054291
940	1.000000	0.999355	-0.161472	-0.165730	0.000000	-0.123388	-0.077092	-0.035115	-0.988122	-0.052575
943	0.999952	0.567314	0.464889	0.286313	0.000000	0.747739	-0.384586	-0.032558	-1.038373	-0.038238
952	0.994324	0.597918	-0.618893	-0.810819	0.000000	0.304285	0.897282	-0.032926	1.411604	-0.054334
963	0.983792	0.961052	-0.199552	-0.043518	0.191935	-0.193316	0.440387	-0.033726	0.007759	-0.054195
965	0.937113	0.992417	-0.054304	-0.145398	0.347112	0.970660	-0.387735	-0.032240	-0.831442	-0.052623
975	0.974463	0.863146	0.624885	0.560262	0.000000	-0.146949	0.655569	-0.031137	-0.476219	-0.053572
986	0.999285	0.998506	0.827404	0.809625	0.956735	0.429413	0.370118	-0.031753	-0.844550	-0.052205
1024	0.996287	0.943707	0.637786	0.435098	-0.598050	0.853106	0.445108	-0.030675	-0.441364	-0.053432
1063	1.000000	0.991599	0.844219	0.823941	0.000000	-0.081792	-0.190562	-0.028441	-1.045474	-0.009374
1188	0.601036	0.820630	-0.193691	-0.591339	0.000000	0.663506	0.240296	-0.032305	-0.964823	-0.048718

Since we considered small clusters to be of size less than 50 points, we didn't get any outliers here as none of the 6 clusters have data points less than 50.

- Finally we get the results of outliers as:

Number of Small Cluster Points: 0

Number of Boundary Points: 62

Total number of outliers: 62