DATA MINING
Input Features Classification Selection Problem

CASE STUDY

Abstract

The case study involves problem statement, methodology, results, analysis and conclusions regarding features selection during classification problems.

SUBMITTED TO: **DR. SAMAN RIAZ**

GROUP MEMBERS

MUHAMMAD FAIZAN ABID (098)

MUHAMMAD USMAN (101)

ZEESHAN MEHDI (007)

SYED MUHAMMAD AMMAR (120)

WAHAJ ALI (FALL-15 / 053 / BSEE)

Table of Contents

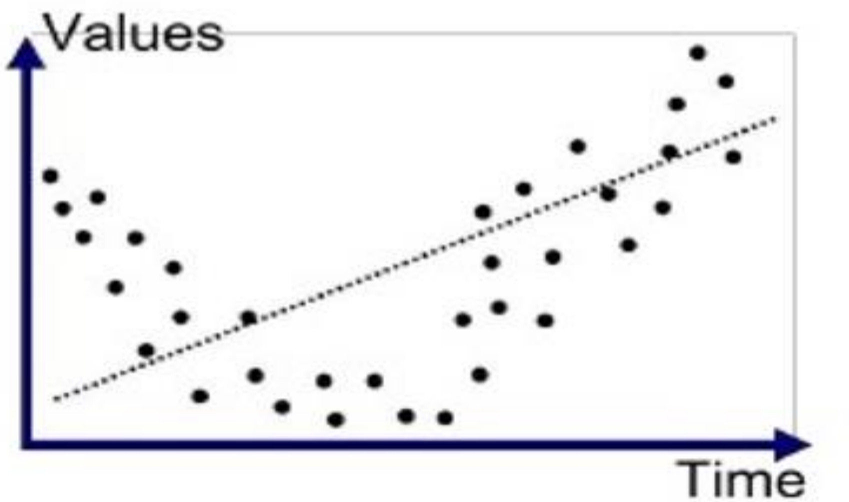
Problem

In this growing era of technology, the data itself is making a high significance in every field either it is business, medicine, stock market, education etc. With growing technology, data itself is growing on a very mass scale which can be a problem for data analyst or data scientist as it is very difficult to manipulate "***high dimensionality***" data-set.

The core problems arsis with high dimensionality data-set are:

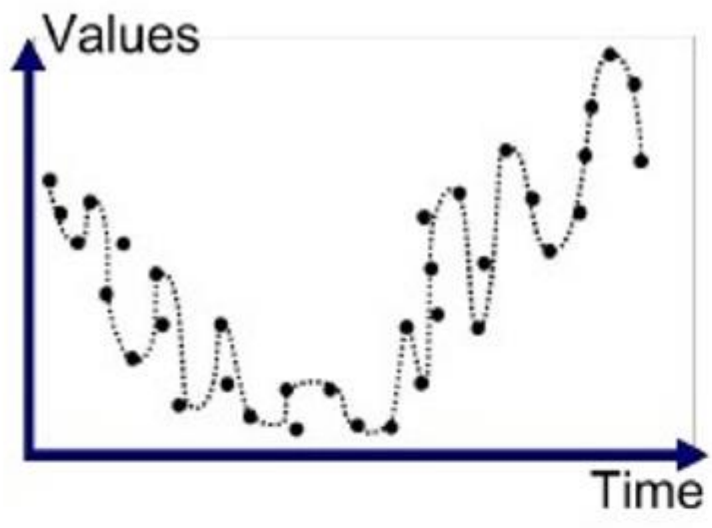
- **Under-Fitting**
    - If a data analyst tries to reduce the dimensions of the data without using any proper technique, by using his own respective sense to related field of data, this can be result in under-fitting as the model

train on very a smaller number of features will not cover most of the data points.



Underfitted

- **Over-Fitting**
    - It is a situation where the respective data analyst tries to cover up all features (including almost every feature), but somehow the train model tried to cover every data point which decrease the accuracy of overall model as the model is confused among the importance of features.



Overfitted

- **Complexity**
    - In the process of Extraction Transformation Loading (ETL), it is commonly seen that data-set with high number of attributes increase the load on machine while pre-processing. Also, once a model is built, the model itself holds too much complexity to generate a result due to high number of attributes.

- **Inconsistency & Redundancy:**
    - Most of the times, we ignore feature engineering (which one to include and which one to exclude) which may cause inaccurate results by trained model as some features contains data inconsistency or redundant data which effects accuracy of model. It is important to highlight such features and to manipulate them respectively them.

- **Time Consumption:**

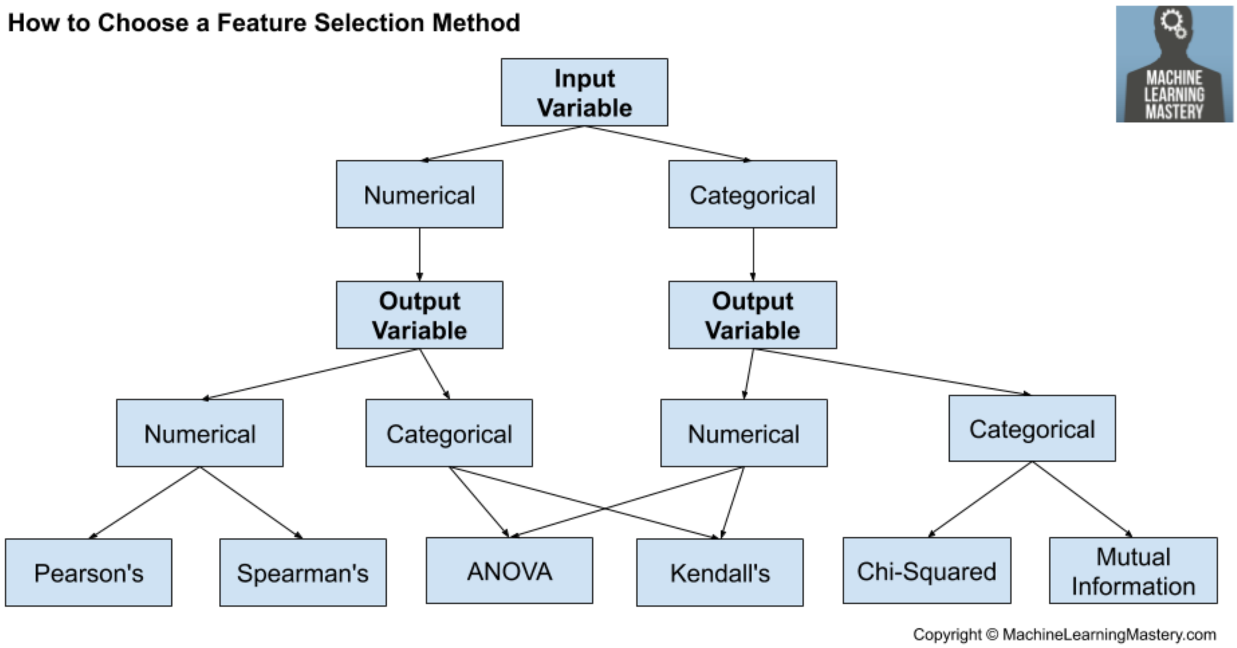
    - Data-sets with high number of dimensions are used to exert load on machine while preprocessing and model training which ultimately cost time consumption and reduce overall time efficiency of the system. So, it is highly necessary to reduce number of features in a rightful manner in order to keep system optimized.

- **Data Representation:**

    - It is nearly impossible to represent data with large number of features. Suppose a data with 50 number of features, you have to make at least 10 different graphs in order to cover all features. In today's world, data representation is very important to create insights in order to grow business.

Solution

In order to tackle the problem of high dimensionality dataset, it is proposed to use some techniques with respect to type of your data in order to reduce dimensionality.



So multiple approaches can be used for dimensionality reduction. However, in this case study we will be using "***Principal Component Analysis (PCA)***" and "***Chi-Square Test (x2)***".

PCA

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

Importantly, the dataset on which PCA technique is to be used must be scaled. The results are also sensitive to the relative scaling. As a layman, it is a method of summarizing data. Imagine some juice bottles on a dining table. Each juice bottle is described by its attributes like color, strength, expiry, etc. But redundancy will arise because many of them will measure related properties. So, what PCA will do in this case is summarize each juice bottle in the stock with less characteristics.

Chi-Square Test

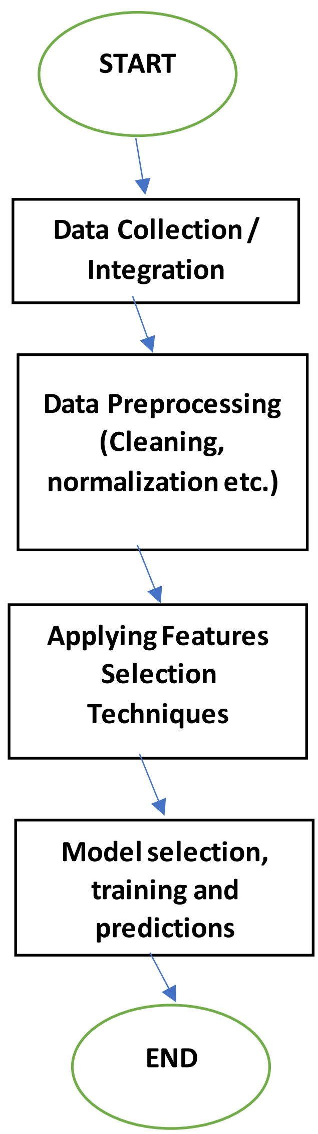
A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

Let's consider a scenario where we need to determine the relationship between the independent category feature (predictor) and dependent category feature(response). In feature selection, we aim to select the features which are highly dependent on the response.

When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

Methodology

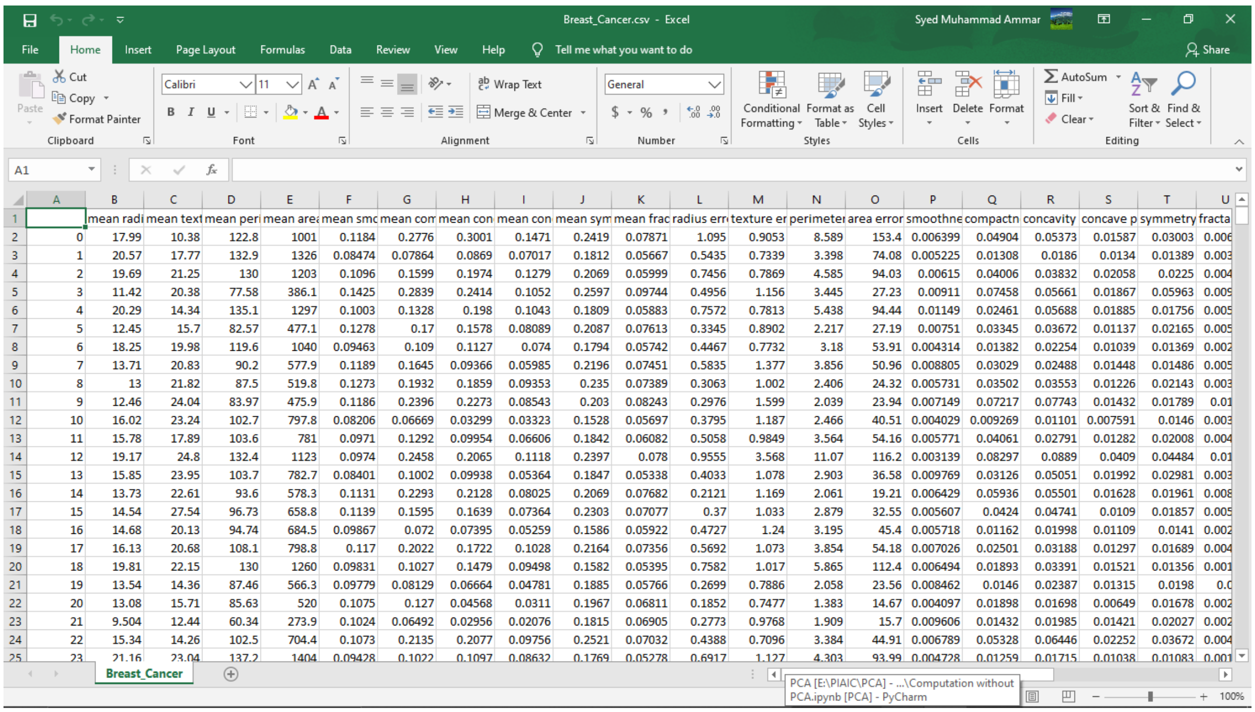
Flow-chart:



Numeric-Dataset:

Our first data-set is of Breast Cancer.

(Snippet of Data-set)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		mean radi	mean text	mean peri	mean area	mean smo	mean com	mean con	mean con	mean sym	mean frac	radius err	texture er	perimeter	area error	smoothne	compactn	concavity	concave p	symmetry	fracta
2	0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006
3	1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003
4	2	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004
5	3	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009
6	4	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005
7	5	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005
8	6	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002
9	7	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005
10	8	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003
11	9	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01
12	10	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003
13	11	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004
14	12	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01
15	13	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003
16	14	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008
17	15	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005
18	16	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002
19	17	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004
20	18	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001
21	19	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0
22	20	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002
23	21	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002
24	22	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03672	0.004
25	23	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01083	0.001

Total # of Rows: 569

Total # of columns: 31 (without target class)

Names of features (all features hold continuous data):

1. mean radius
2. mean radius
3. mean texture

4. mean perimeter
5. mean area
6. mean smoothness
7. mean compactness
8. mean concavity
9. mean concave points
10. mean symmetry
11. mean fractal dimension
12. radius error
13. texture error
14. perimeter error
15. area error
16. smoothness error
17. compactness error
18. concavity error
19. concave points error
20. symmetry error
21. fractal dimension error
22. worst radius
23. worst texture
24. worst perimeter
25. worst area
26. worst smoothness
27. worst compactness
28. worst concavity
29. worst concave points
30. worst symmetry
31. worst fractal dimension

Target Class: Is_Cancer

Target State: Binary (0 or 1)

Preprocessing

For every model training, pre-processing steps are very important. Coming towards this data-set, this data-set holds all numeric values (continuous data). So, in order to preprocess this data, there are multiple steps.

First of all, we will cater null values if there any by using mean method. Also, different methods can also be used to fill null values.

So now, after this step our data looks clean enough to proceed. Now coming toward most important step is “dimensionality reduction”. By looking at this data, we can suggest PCA algorithm as it works best on the continuous data.

PCA steps: transform an $N \times d$ matrix X into an $N \times m$ matrix Y:

- ❑Centralized the data (subtract the mean).
- ❑Calculate the $d \times d$ covariance matrix: $C = \frac{1}{N-1} X^T X$
  - ❑ $C_{i,j} = \frac{1}{N-1} \sum_{q=1}^{N} X_{q,i} . X_{q,j}$
  - ❑ $C_{i,i}$ (diagonal) is the variance of variable i.
  - ❑ $C_{i,j}$ (off-diagonal) is the covariance between variables i and j.
- ❑Calculate the eigenvectors of the covariance matrix (orthonormal).

As our current data-set holds 31 features (mentioned in Result & Analysis), we will convert them into 5 new features which will hold most of the information. One thing very important here to standardization of the data as it effects PCA performance.

Here is the snippet of data before standardization.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	w smoothn
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	25.380	17.33	184.60	2019.0	0.16
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	24.990	23.41	158.80	1956.0	0.12
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	23.570	25.53	152.50	1709.0	0.14
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	14.910	26.50	98.87	567.7	0.20
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	22.540	16.67	152.20	1575.0	0.13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	25.450	26.40	166.10	2027.0	0.14
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	23.690	38.25	155.00	1731.0	0.11
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	18.980	34.12	126.70	1124.0	0.11
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	25.740	39.42	184.60	1821.0	0.16
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	9.456	30.37	59.16	268.6	0.08

569 rows × 30 columns

Here is snippet of data after standardization

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter
0	1.097064	-2.073335	1.269934	0.984375	1.568466	3.283515	2.652874	2.532475	2.217515	2.255747	...	1.886690	-1.359293	2.303601
1	1.829821	-0.353632	1.685955	1.908708	-0.826962	-0.487072	-0.023846	0.548144	0.001392	-0.868652	...	1.805927	-0.369203	1.535126
2	1.579888	0.456187	1.566503	1.558884	0.942210	1.052926	1.363478	2.037231	0.939685	-0.398008	...	1.511870	-0.023974	1.347475
3	-0.768909	0.253732	-0.592687	-0.764464	3.283553	3.402909	1.915897	1.451707	2.867383	4.910919	...	-0.281464	0.133984	-0.249939
4	1.750297	-1.151816	1.776573	1.826229	0.280372	0.539340	1.371011	1.428493	-0.009560	-0.562450	...	1.298575	-1.466770	1.338539
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	2.110995	0.721473	2.060786	2.343856	1.041842	0.219060	1.947285	2.320965	-0.312589	-0.931027	...	1.901185	0.117700	1.752563
565	1.704854	2.085134	1.615931	1.723842	0.102458	-0.017833	0.693043	1.263669	-0.217664	-1.058611	...	1.536720	2.047399	1.421940
566	0.702284	2.045574	0.672676	0.577953	-0.840484	-0.038680	0.046588	0.105777	-0.809117	-0.895587	...	0.561361	1.374854	0.579001
567	1.838341	2.336457	1.982524	1.735218	1.525767	3.272144	3.296944	2.658866	2.137194	1.043695	...	1.961239	2.237926	2.303601
568	-1.808401	1.221792	-1.814389	-1.347789	-3.112085	-1.150752	-1.114873	-1.261820	-0.820070	-0.561032	...	-1.410893	0.764190	-1.432735

569 rows × 30 columns

So now we split data into training & testing data and perform PCA separately. After implementing PCA, we get data-set with 5 PCA-Components.

Training Data:

	PCA-1	PCA-2	PCA-3	PCA-4	PCA-5
0	-2.158330	1.698969	-1.095885	-1.163401	-0.567506
1	3.762117	0.931613	3.734847	-1.581371	-2.467616
2	-2.213925	-1.836971	-0.238475	-0.974115	0.584291
3	1.599827	2.203594	-3.156443	-0.169390	-1.109276
4	1.650250	1.459177	-1.917315	1.284165	-0.939887
...	...	...	...	...	...
450	-1.248698	0.769058	0.899252	4.002755	-0.864534
451	-4.551813	-2.781786	1.164885	-0.194569	0.239992
452	-3.221108	-2.247268	0.041821	-1.162988	1.304741
453	-4.643109	-0.284077	1.652673	-0.107984	-2.034921
454	12.921333	2.684580	6.269831	-1.404877	-3.142794

455 rows × 5 columns

Testing Data:

	PCA-1	PCA-2	PCA-3	PCA-4	PCA-5
0	0.525166	0.321703	-0.952458	-0.433863	1.022118
1	-1.736773	0.782417	2.475560	1.275663	2.308672
2	6.570045	-1.911929	0.005279	-0.713947	-0.484667
3	3.075021	-1.368350	2.682932	-1.033881	-0.536782
4	-0.839703	-2.074784	-0.211020	-0.214358	-0.444422
...	...	...	...	...	...
109	8.900092	-1.302009	1.622525	1.010093	-0.023199
110	0.988026	1.011470	0.580263	1.960171	0.588477
111	4.398378	6.175283	-3.050683	3.228339	0.270585
112	2.107139	0.936916	1.443023	-0.978948	-1.190131
113	-1.562892	1.190187	-1.807731	1.332132	-0.341807

114 rows × 5 columns

So now we are good to proceed towards model fitting. As this is classification problem so we will be using three different models.

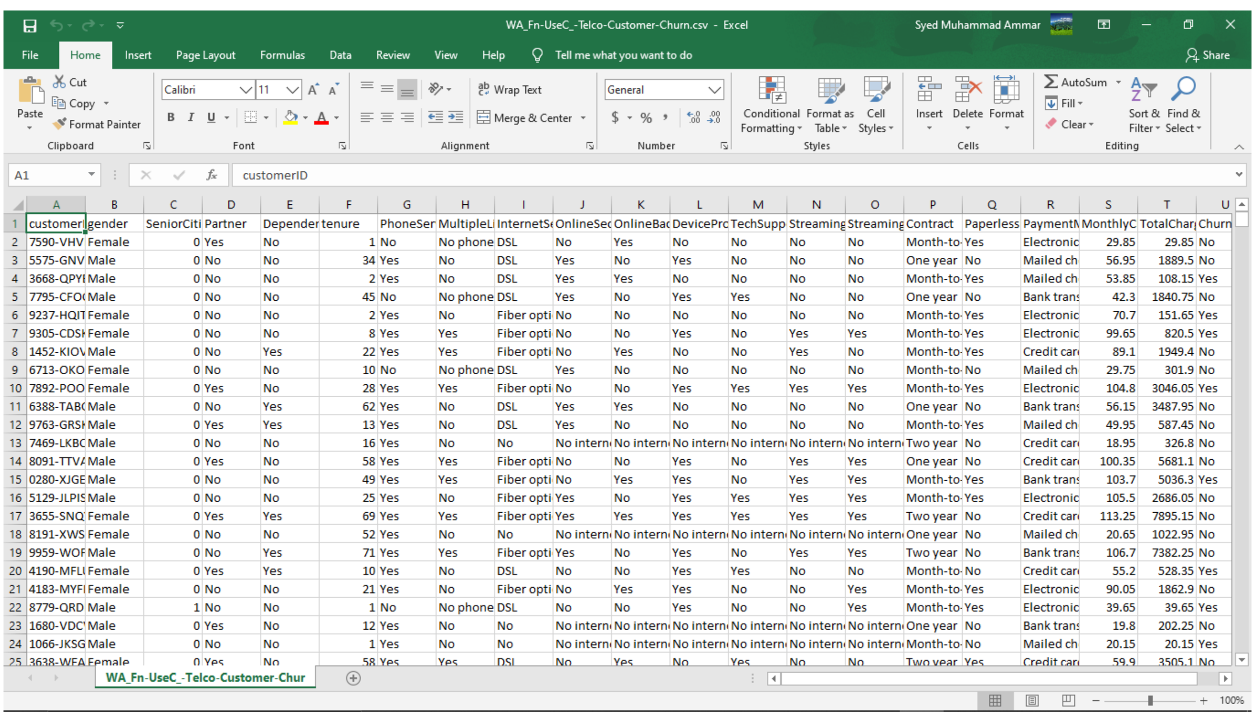
1. Decision Tree
2. Logistic Regression

The performance is discussed in Result and Analysis part.

Categorical Dataset:

Our second data-set is of Telco Customer.

(Snippet of Data-set)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerI	gender	SeniorCiti	Partner	Depender	tenure	PhoneSer	MultipleLi	InternetS	OnlineSec	OnlineBac	DevicePrc	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentM	MonthlyC	TotalChar	Churn
2	7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed ch	56.95	1889.5	No
4	3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed ch	53.85	108.15	Yes
5	7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
6	9237-HQIT	Female	0	No	No	2	Yes	No	Fiber opti	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes
7	9305-CDSK	Female	0	No	No	8	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit car	89.1	1949.4	No
9	6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed ch	29.75	301.9	No
10	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber opti	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes
11	6388-TABC	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
12	9763-GRSK	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed ch	49.95	587.45	No
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Two year	No	Credit car	18.95	326.8	No
14	8091-TTVA	Male	0	Yes	No	58	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
15	0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber opti	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber opti	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No
17	3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
18	8191-XWS	Female	0	No	No	52	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Mailed ch	20.65	1022.95	No
19	9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber opti	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
20	4190-MFLU	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit car	55.2	528.35	Yes
21	4183-MYFI	Female	0	No	No	21	Yes	No	Fiber opti	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No
22	8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic	39.65	39.65	Yes
23	1680-VDC	Male	0	Yes	No	12	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Bank trans	19.8	202.25	No
24	1066-JKSG	Male	0	No	No	1	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Month-to-	No	Mailed ch	20.15	20.15	Yes
25	3638-WFA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit car	59.9	3505.1	No

Total # of Rows: 7043

Total # of columns: 20 (without target class)

Names of features:

1. customerID (categorical)
2. gender (categorical)
3. SeniorCitizen (categorical)
4. Partner (categorical)
5. Dependents (categorical)
6. Tenure (integer)
7. PhoneService (categorical)
8. MultipleLines (categorical)
9. InternetService (categorical)
10. OnlineSecurity (categorical)
11. OnlineBackup (categorical)
12. DeviceProtection (categorical)
13. TechSupport (categorical)
14. StreamingTV (categorical)
15. StreamingMovies (categorical)
16. Contract (categorical)
17. PaperlessBilling (categorical)
18. PaymentMethod (categorical)
19. MonthlyCharges (continuous)
20. TotalCharges (continuous)

Target Class: Churn (categorical)

Target State: Binary (Yes or No)

Preprocessing

For every model training, pre-processing steps are very important. Coming towards this data-set, this data-set holds most of features as categorical and a few as numerical or continuous. So, in order to preprocess this data, there are multiple steps.

First of all, we remove customerID, numerical and continuous features from data to make data fully categorical.

After that, we cater null values if there any by using mode method. Also, different methods can also be used to fill null values.

gender	SeniorCitizen	Partner	Dependents	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	Strea
Female	0	Yes	No	No	No phone service	DSL	No	Yes	No	No	
Male	0	No	No	Yes	No	DSL	Yes	No	Yes	No	
Male	0	No	No	Yes	No	DSL	Yes	Yes	No	No	
Male	0	No	No	No	No phone service	DSL	Yes	No	Yes	Yes	
Female	0	No	No	Yes	No	Fiber optic	No	No	No	No	
...	...	...	...	...	...	...	...	...	...	...	
Male	0	Yes	Yes	Yes	Yes	DSL	Yes	No	Yes	Yes	
Female	0	Yes	Yes	Yes	Yes	Fiber optic	No	Yes	Yes	No	
Female	0	Yes	Yes	No	No phone service	DSL	Yes	No	No	No	
Male	1	Yes	No	Yes	Yes	Fiber optic	No	No	No	No	
Male	0	No	No	Yes	No	Fiber optic	Yes	No	Yes	Yes	

So now, after this step our data looks clean enough to proceed. Now coming toward most important step is "dimensionality reduction". By looking at this data, we can suggest Chi-Square Test as it works best on the categorical data.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

χ^2 = Chi Square obtained
$\sum$ = the sum of
O = observed score
E = expected score

Let's take a view of result of chi-square test which tell us about features importance. The features are arranged based on p-values (ascending order).

FEATURES	P-VALUES
TechSupport	1.186565e-95
OnlineSecurity	1.619912e-95
Contract	1.127358e-82
InternetService	3.154657e-62
OnlineBackup	1.537075e-50
DeviceProtection	6.000278e-36
Dependents	3.870061e-26
SeniorCitizen	2.817944e-25
PaperlessBilling	8.276605e-20

Partner	6.505617e-16
PaymentMethod	1.928091e-10
StreamingMovies	9.883292e-04
MultipleLines	2.621376e-03
StreamingTV	2.556954e-02
gender	4.304357e-01
PhoneService	6.710351e-01

So, we can clearly see that the most important feature is “Tech Support”. Now we finalized our final data-set by picking up top 8 features.

The dataset is then split into training and testing data. Random forest classifier is applied to gain the predictions (results).

The performance is discussed in Result and Analysis part.

Result & Analysis

In this part of case study, we will be covering the main aspects with respect to performance. The analysis will be done with respect to time complexity of model and confusion matrix of model.

Numeric Dataset Results & Analysis

TRAINING TIME

Models	Training Time with PCA	Training Time without PCA	Gain With PCA
Decision Tree	3.99ms	13ms	69.31%
Logistic Regression	7.99ms	75.6ms	89.44%

PREDICTION TIME

Models	Prediction Time with PCA	Prediction Time without PCA	Gain with PCA
Decision Tree	997µs	998µs	0.11%
Logistic Regression	997µs	1.01ms	1.29%

SIMPLE ACCURACY

Models	Simple Accuracy with PCA	Simple Accuracy without PCA	Gain with PCA
Decision Tree	0.9	0.89	1.12%
Logistic Regression	0.92	0.86	6.53%

PRECISION SCORE

Models	Precision Score with PCA	Precision Score without PCA	Gain with PCA
Decision Tree	0.87	0.85	2.3%
Logistic Regression	0.9	0.87	3.4%

RECALL SCORE

Models	Recall Score with PCA	Recall Score without PCA	Gain with PCA
Decision Tree	1	1	0%
Logistic Regression	1	0.93	7%

F1-SCORE

Models	F1-Score with PCA	F1-Score without PCA	Gain with PCA
Decision Tree	0.94	0.91	3.2%
Logistic Regression	0.95	0.9	5.27%

Categorical Dataset Results & Analysis

TRAINING TIME

Models	Training Time with Chi-Square Test	Training Time without Chi-Square Test	Gain with Chi-Square Test
Random Forest	404ms	1.1s	63.28%

SIMPLE ACCURACY

Models	Simple Accuracy with Chi-Square Test	Simple Accuracy without Chi-Square Test	Gain with Chi-Square Test
Random Forest	0.81	0.7	14%

PRECISION SCORE

Models	Precision Score with Chi-Square Test	Precision Score without Chi-Square Test	Gain with Chi-Square Test
Random Forest	0.61	0.42	31.15%

RECALL SCORE

Models	Recall Score with Chi-Square Test	Recall Score without Chi-Square Test	Gain with Chi-Square Test
Random Forest	0.6	0.52	16.13%

F1-SCORE

Models	F1-Score with Chi-Square Test	F1-Score without Chi-Square Test	Gain with Chi-Square Test
Random Forest	0.61	0.46	24.6%

CONCLUSION

From the case study, we came to conclusion that feature selection is a very important and effective step in data-preprocessing as it boosts your methodology in terms of time efficiency, model accuracy and data representation. There are multiple feature selection techniques which can be used with respect to type of data you are treating.