SDSC3006 FUNDAMENTALS OF MACHINE LEARNING I PROJECT

Tittle: Comparing Logistic Regression Models with and without Principal Component

DATE: 2023-11-25

TSUI NGA CHING 56621075
LAM LOK HIN 56474716
FUXIAN ZHAO 40149034

TEAM MEMBER DISTRIBUTION

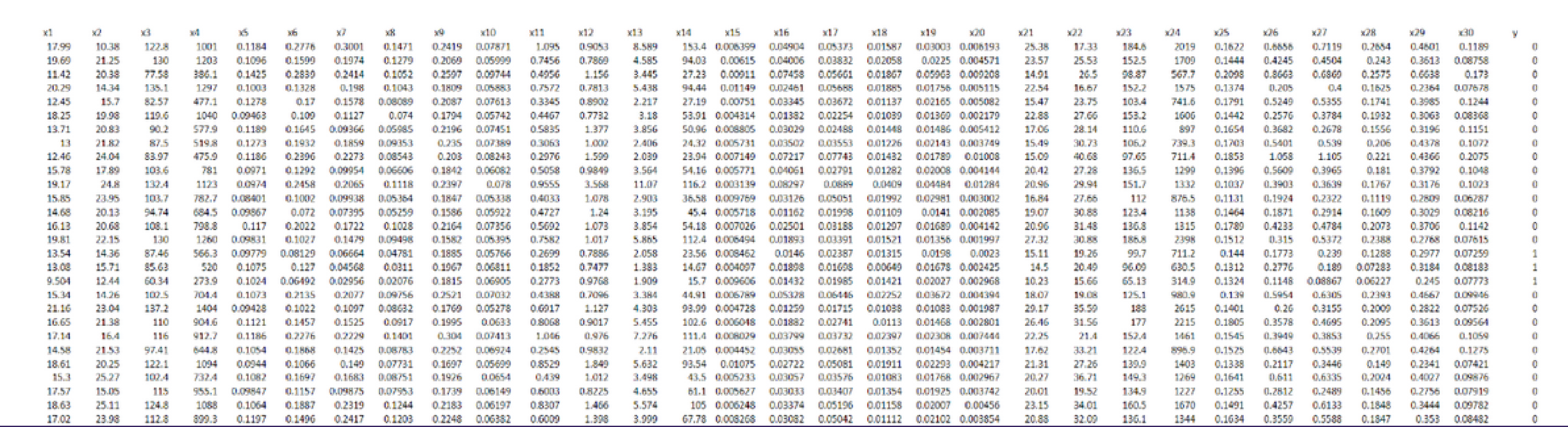
Tsui Nga Ching: Info, Problem Formulation, Strategies and Methods, Conclusion & discussion

Lam Lok Hin :Logistic Regression- justification, data analysis, results

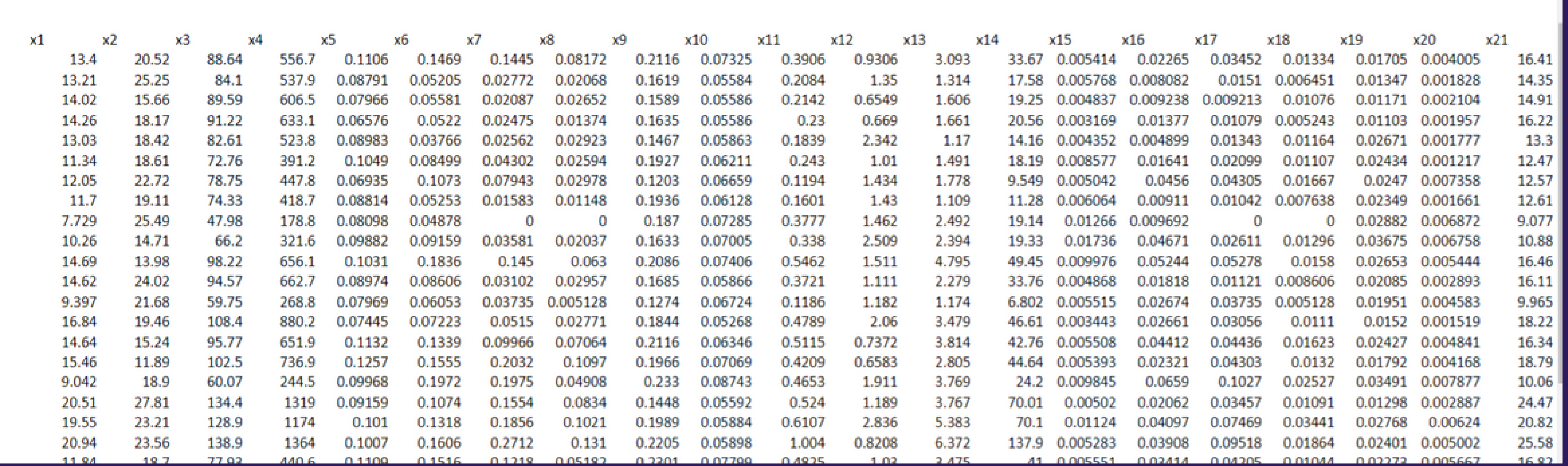
Fuxian Zhao:PCA, try LDA,SVM-Justification, results

1. INFO

- Two datasets: **Training.xlsx (n = 519)** and **test.xlsx (n = 50)**
- Variables: 30 predictors (X1, X2,...,X30), all numerical
- 1 response variable (Y) with two classes (0 and 1)
- Test data has missing values for the response variable



TRAINING



x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21
13.4	20.52	88.64	556.7	0.1106	0.1469	0.1445	0.08172	0.2116	0.07325	0.3906	0.9306	3.093	33.67	0.005414	0.02265	0.03452	0.01334	0.01705	0.004005	16.41
13.21	25.25	84.1	537.9	0.08791	0.05205	0.02772	0.02068	0.1619	0.05584	0.2084	1.35	1.314	17.58	0.005768	0.008082	0.0151	0.006451	0.01347	0.001828	14.35
14.02	15.66	89.59	606.5	0.07966	0.05581	0.02087	0.02652	0.1589	0.05586	0.2142	0.6549	1.606	19.25	0.004837	0.009238	0.009213	0.01076	0.01171	0.002104	14.91
14.26	18.17	91.22	633.1	0.06576	0.0522	0.02475	0.01374	0.1635	0.05586	0.23	0.669	1.661	20.56	0.003169	0.01377	0.01079	0.005243	0.01103	0.001957	16.22
13.03	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.02923	0.1467	0.05863	0.1839	2.342	1.17	14.16	0.004352	0.004899	0.01343	0.01164	0.02671	0.001777	13.3
11.34	18.61	72.76	391.2	0.1049	0.08499	0.04302	0.02594	0.1927	0.06211	0.243	1.01	1.491	18.19	0.008577	0.01641	0.02099	0.01107	0.02434	0.001217	12.47
12.05	22.72	78.75	447.8	0.06935	0.1073	0.07943	0.02978	0.1203	0.06659	0.1194	1.434	1.778	9.549	0.005042	0.0456	0.04305	0.01667	0.0247	0.007358	12.57
11.7	19.11	74.33	418.7	0.08814	0.05253	0.01583	0.01148	0.1936	0.06128	0.1601	1.43	1.109	11.28	0.006064	0.00911	0.01042	0.007638	0.02349	0.001661	12.61
7.729	25.49	47.98	178.8	0.08098	0.04878	0	0	0.187	0.07285	0.3777	1.462	2.492	19.14	0.01266	0.009692	0	0	0.02882	0.006872	9.077
10.26	14.71	66.2	321.6	0.09882	0.09159	0.03581	0.02037	0.1633	0.07005	0.338	2.509	2.394	19.33	0.01736	0.04671	0.02611	0.01296	0.03675	0.006758	10.88
14.69	13.98	98.22	656.1	0.1031	0.1836	0.145	0.063	0.2086	0.07406	0.5462	1.511	4.795	49.45	0.009976	0.05244	0.05278	0.0158	0.02653	0.005444	16.46
14.62	24.02	94.57	662.7	0.08974	0.08606	0.03102	0.02957	0.1685	0.05866	0.3721	1.111	2.279	33.76	0.004868	0.01818	0.01121	0.008606	0.02085	0.002893	16.11
9.397	21.68	59.75	268.8	0.07969	0.06053	0.03735	0.005128	0.1274	0.06724	0.1186	1.182	1.174	6.802	0.005515	0.02674	0.03735	0.005128	0.01951	0.004583	9.965
16.84	19.46	108.4	880.2	0.07445	0.07223	0.0515	0.02771	0.1844	0.05268	0.4789	2.06	3.479	46.61	0.003443	0.02661	0.03056	0.0111	0.0152	0.001519	18.22
14.64	15.24	95.77	651.9	0.1132	0.1339	0.09966	0.07064	0.2116	0.06346	0.5115	0.7372	3.814	42.76	0.005508	0.04412	0.04436	0.01623	0.02427	0.004841	16.34
15.46	11.89	102.5	736.9	0.1257	0.1555	0.2032	0.1097	0.1966	0.07069	0.4209	0.6583	2.805	44.64	0.005393	0.02321	0.04303	0.0132	0.01792	0.004168	18.79
9.042	18.9	60.07	244.5	0.09968	0.1972	0.1975	0.04908	0.233	0.08743	0.4653	1.911	3.769	24.2	0.009845	0.0659	0.1027	0.02527	0.03491	0.007877	10.06
20.51	27.81	134.4	1319	0.09159	0.1074	0.1554	0.0834	0.1448	0.05592	0.524	1.189	3.767	70.01	0.00502	0.02062	0.03457	0.01091	0.01298	0.002887	24.47
19.55	23.21	128.9	1174	0.101	0.1318	0.1856	0.1021	0.1989	0.05884	0.6107	2.836	5.383	70.1	0.01124	0.04097	0.07469	0.03441	0.02768	0.00624	20.82
20.94	23.56	138.9	1364	0.1007	0.1606	0.2712	0.131	0.2205	0.05898	1.004	0.8208	6.372	137.9	0.005283	0.03908	0.09518	0.01864	0.02401	0.005002	25.58
11.84	18.7	77.93	440.6	0.1109	0.1516	0.1218	0.05182	0.2301	0.07799	0.4825	1.03	3.475	41	0.005551	0.03414	0.04205	0.01044	0.02273	0.005667	16.82

TESTING

2. PROBLEM FORMULATION

- The problem at hand is to apply the knowledge and skills acquired in data analysis to analyze real data.
- The objective is to interpret the results of the data analysis and present the findings.
- Specifically, we apply **LDA,SVM,Logistic Regression** to the training dataset for prediction and decide **Logistic Regression** as
- our final model because of its smallest overfitting problem.
- To solve overfitting problem further ,we use **PCA** to reduce dimension of data.
- Programming Language used: **Python**

3. STRATEGIES AND METHODS

Strategy:

The strategy is to compare **accuracy** and **sensltivity** (Recall) between different models and whether the same model uses PCA accoring to cross validation and find the best model.

Methods:

Logistic regression: a statistical model used to predict binary outcomes by fitting a logistic function to the observed data.

Linear discriminant analysis: a supervised dimensionality reduction technique that optimally transforms input features to maximize the separation between classes while minimizing within-class variance

SVM(linear kernel): a supervised machine learning algorithm that separates classes in a dataset by finding the hyperplane that maximally separates the support vectors, representing instances near the class boundaries.

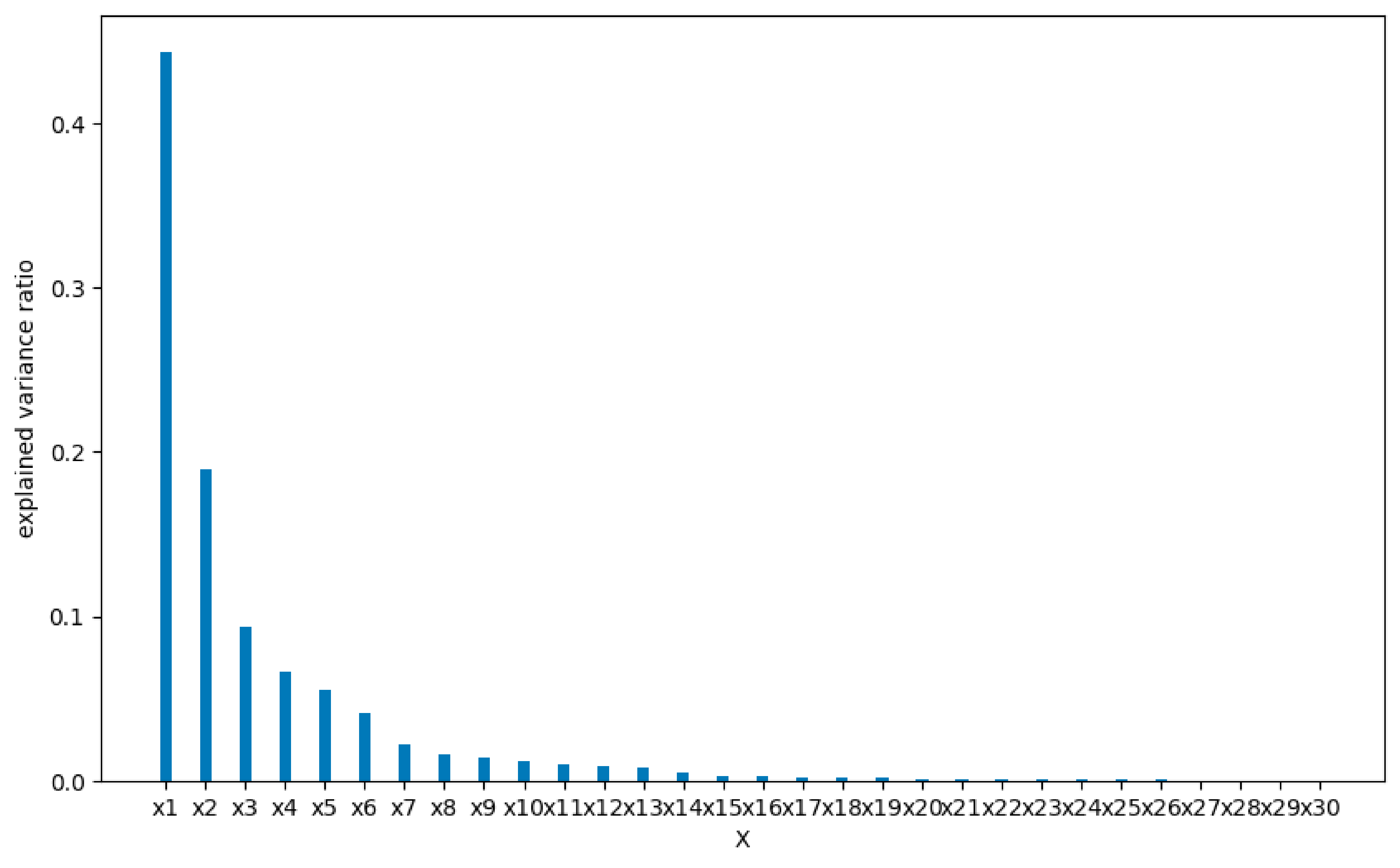
4.JUSTIFICATION

1) Data Processing

- import data and relevant package
- check null values and outliers

2)Principal Component Analysis

- Standardizes the training data
- Fits a PCA model to the standardized data
- visualization



5.DATA ANALYSIS

After the data preparation, the training data will be fitted in the models. The accuracy and sensitivity of cross validation are used to measure model performance and the accuracy on training data is used to measure overfitting level.For each model ,we use original data and data with top 8 principal variables.

6. RESULTS

MODEL	FITTING ON DATA	ACCURACY	RECALL	ACCURACY ON TRAINING DATA
LDA	ORIGINAL DATA	0.9262574257425742	0.9814903846153846	0.940
	PRINCIPAL DATA	0.9268297236743838	0.9723543123543124	0.933
SVM(LINEAR KERNEL)	ORIGINAL DATA(C=25)	0.9560990099009901	0.9722596153846155	0.964
	PRINCIPAL DATA(C=15)	0.9162178217821783	0.953798076923077	0.924
LOGISTIC REGRESSION	ORIGINAL DATA	0.9422516803584763	0.9633100233100234	0.946
	PRINCIPAL DATA	0.9042376237623764	0.947644230769231	0.915

Considering accuracy and recall, SVM fitted on the original data performs best, but there is overfitting problem.So,Logistic regression fitted on data with top8 principal variables perform best which has lowest accuracy on training data.

7. CONCLUSION& DISCUSSION

Conclusion:

- Considering overfitting level ,the logistic regression with top 8 principal variables is selected as the final model.According to cross validation, the estimated accuracy and sensitivity are 90.42% and 94.76%,which means most positive data are predicted correctly with high accuracy.

7. CONCLUSION& DISCUSSION

limitation

- Although we try different models and use PCA to reduce dimensions, we don't solve overfitting problem because the accuracy on training data still high.
- However, logistic regression are relevantly simple model, which doesn't cause overfitting problem easily. So, we think the effective way to avoid this problem is to increase amount of data.
- For overfitting model, the estimation of cross validation is not reliable enough.So, we don't have enough confidence to measure the model performance on test data.