Dimensionality Reduction: Principal Component Analysis

Chris Haddad, Jeff Coady, Sanjay Roberts

COMP 3441 - University of Denver

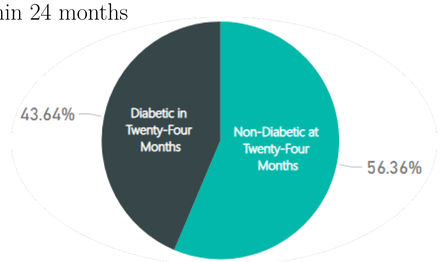
Objective

Use Principal Component Analysis to reduce dimensionality and model execution time, relative to using all predictors, while maintaining prediction accuracy.

Dataset

Allscripts Pre-Diabetic to Diabetic Transition Data

- 150,000 pre-diabetic patients tracked over time.
- 44% develop diabetes
- Tracked factors predict transition to type II diabetes within 24 months



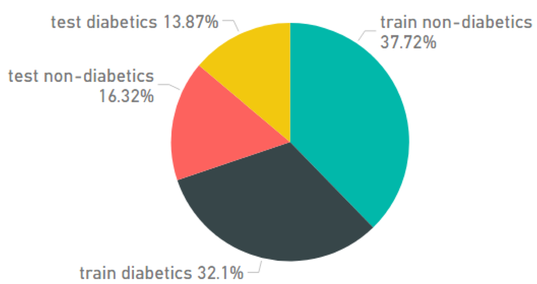
Features

Feature Columns		
Blood Pressure	BMI	Age
Ethnicity	Race	Gender
Family History	HbA1c	

Experiment

Random Forest Ensemble Learning method as classification predictor

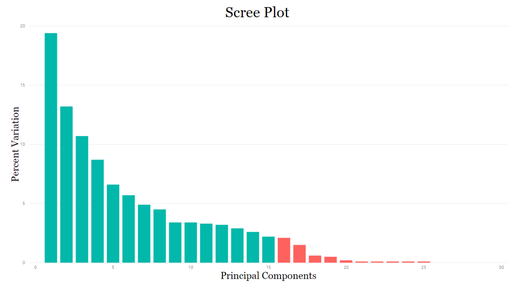
- Train model without PCA
- Record time and accuracy
- Replicate method using PCs that describe 95% of variance and less

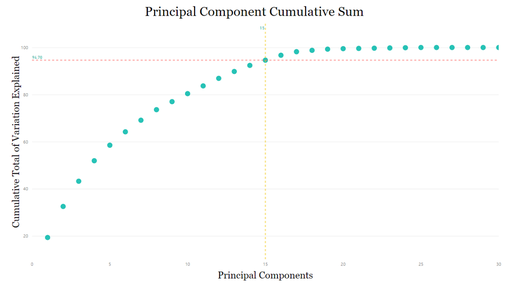


PCA

Takes correlated predictor variables and creates set of uncorrelated components to be used instead.

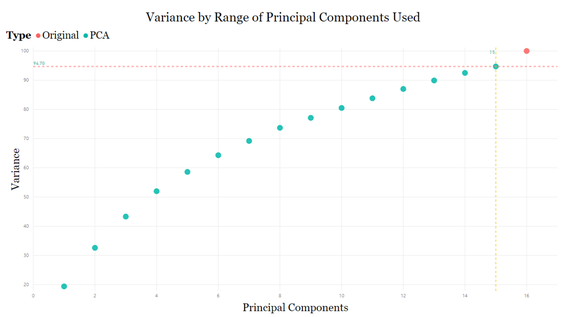
1. Center and scale data
2. Calculate eigenvectors and eigenvalues
3. Principal components determined by which predictors contribute most to an eigenvector
4. Subset of PCs chosen in place of original





Variance

- Model uses 15 of 30 components which account for ≈ 95% of variance

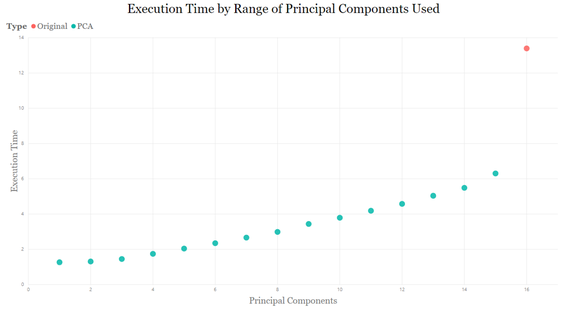


Factor Loadings

Factor	PC 1	PC 2	PC 3	PC 4	PC 5
Sys Avg	-.298	-.169	-.216	.009	.066
Sys Min	-.256	-.155	-.180	-.020	.013
Sys Max	-.250	-.139	-.186	.030	.096
Sys Last	-.261	-.158	-.197	.003	.060
Dias Avg	-.284	-.251	-.065	-.094	.030
Dias Min	-.227	-.207	-.059	-.105	-.014
Dias Max	-.249	-.217	-.055	-.058	.066
Dias Last	-.244	-.227	-.065	-.089	.031
BMI Avg	-.284	.096	.379	.042	-.070
BMI Min	-.283	.094	.372	.039	-.074
BMI Max	-.280	.096	.378	.043	-.066
BMI Last	-.283	.094	.377	.041	-.072
HbA1c Avg	-.180	.406	-.196	-.006	.064
HbA1c Min	-.177	.397	-.196	-.010	.070
HbA1c Max	-.177	.401	-.190	-.002	.058
HbA1c Last	-.178	.402	-.195	-.006	.064
Eth Hisp	.026	.032	.052	-.289	-.006
Eth Not Hisp	-.022	-.043	-.035	.528	.110
Eth Unkn	.005	.025	.001	-.397	-.126
White	-.001	-.047	.012	.433	-.106
Unknown	.015	.041	.013	-.435	-.085
Male	-.056	.011	-.165	.052	-.646
Female	.056	-.011	.165	-.052	.646
Hispanic	.001	.026	.013	-.146	.004
Other	.016	-.005	.016	-.112	.023
Black	-.053	.019	.015	-.037	.201
Abnorm BP	-.030	-.057	.003	.046	.030
Family	-.025	.053	.046	-.041	.106
Asian	.051	-.004	-.099	-.028	.068
Age	-.019	.048	-.208	.136	-.007

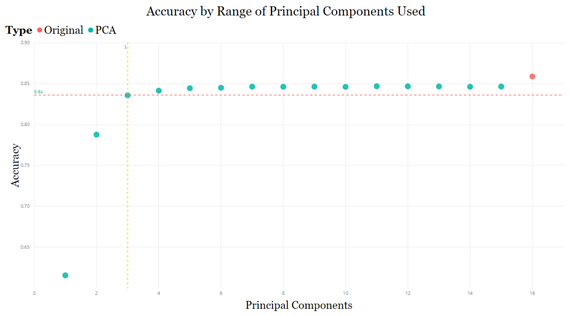
Execution Time

- Without PCA, model takes around 30 minutes to run.

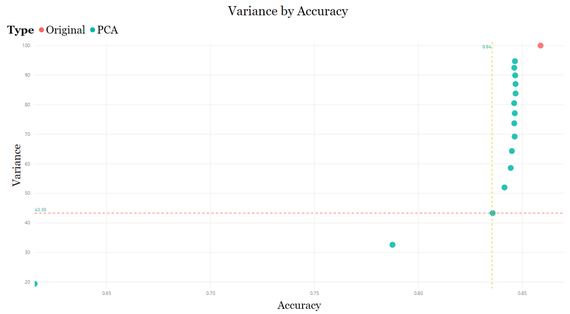


Accuracy

- Model accuracy stable after 3 PCs



- Model accuracy stable after 43% variance



Final Thoughts

- Successfully used PCA to decrease execution time while maintaining accuracy
- Feature factoring loadings split nicely:
  1. Blood Pressure and BMI
  2. HbA1c
  3. BMI
  4. Race and Ethnicity
  5. Gender
- Inflection point at 84% accuracy and 43% total variance accounted for
- ≈ 10% of execution time to reach 84% accuracy as compared to training model on original data
- PCA is an effective method for use in prediction model parameter tunning