|  |
| --- |
| **Machine Learning** |
| **Regression Hackathon** |
| **Authors: Julia Hernández Elena, Federico Soriano Palacios (Team 28)** |
| Data preprocessing:  Primero hemos hecho un análisis de correlaciones y eliminado 10 variables ya que como nos dan valores por posición, las posiciones cercanas están muy relacionadas.  También eliminamos los valores atípicos tras comparar los residuos vs real. Hemos considerado atípicos todos aquellos valores que nos dan un residuo mayor de 50.  Como punto de partida utilizamos un modelo lineal para poder comparar con modelos más complejos. En WD vs WS vimos que la relación lineal no nos valía para predecir los valores de WD ya que obteníamos valores negativos o muy altos que no tenían sentido.  El modelo GAM es muy útil para los casos con una alta dimensión del espacio de entrada, lo utilizamos para ajustar lo que habíamos comprobado en el lineal y obtuvimos mejor RMSE, además con el análisis de este nuevo modelo descubrimos 3 nuevas variables poco relevantes que pudimos eliminar.  Después probamos un modelo MLP, pero al mirar los resultados comprobamos las diferencias entre el entrenamiento y el test, habíamos sobreentrenado el modelo al eliminar los valores atípicos.  Finalmente, para que los valores atípicos no fuesen tan influyentes en el modelo optamos por utilizar un SVM. Cuando conseguimos ajustar este modelo obtuvimos resultados muy parecidos para entrenamiento y para test, nuestro modelo tiene muy buena capacidad de generalización.  Model comparison:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Model | Structure | Inputs | E training | E cross val | E validation | | GAM | df = 13 | Eliminamos outliers y TL2H80, TL3H80, TL4H80, TL5H80, TL6H80, TL7H80, TL8H80, TL10H80, WSL1H80, WSL10H80, WSL7H80 y WSL9H80 | RMSE  15.8679 | RMSE  18.4565 | 22.64 | | GAM | df = 12.5 | Eliminando lo mismo que el anterior mas  WSL3H80, WSL6H80 y WSL8H80 | RMSE  14.78871 | RMSE  17.37615 | 22.53 | | Lineal |  | Eliminamos  TL2H80, TL3H80, TL4H80, TL5H80, TL6H80, TL7H80, TL8H80, TL10H80, WSL1H80, WSL10H80, WSL7H80 y WSL9H80 | RMSE  23.24282 | RMSE  23.42546 | Sin probar | | PLSR | ncomp = 13 |  | R2  0.9215176  RMSE  16.35589 | R2  0.9096651  RMSE  18.33959 | 23.1575 | | Gam | df = 10  14.29116 0.9405365 |  | R2  0.962946  RMSE  11.24278 | R2  0.9428562  RMSE  14.47497 | 21.379 | | MLP |  | Eliminación de valores atípicos | R2  0.992245  RMSE  5.145416 | R2  0.9410525  RMSE  14.85008 | 21.68592 | | SVM | sigma = 0.006812921 C = 7.196857 | Truncamos la salida de forma que las predicciones con valor <0 se ajustan a 0 | R2  0.9403331  RMSE  14.58288 | R2  0.9408054  RMSE  14.6647 | 20.75501 |   Conclusions:  En problemas en los que tengamos un espacio de entrada de altas dimensiones la regularización juega un papel muy importante. La selección de las variables a utilizar en nuestro modelo es vital.  Por otro lado, a la hora de utilizar modelos muy flexibles como el MLP hay que tener mucho cuidado con aumentar la complejidad del modelo sin aumentar el tamaño del conjunto de entrenamiento, ya que al hacer esto se corre peligro de sobreentrenamiento. Por eso mismo obtuvimos unos valores tan buenos para el conjunto de test cuando quitamos los valores atípicos y unos valores tan distintos para validación.  En esos casos, es preferible utilizar modelos como el SVM, no tan sensibles a valores atípicos y que por tanto producen mejores resultados. |
|  |