Práctica 1

Competición en Kaggle sobre Clasificación Binaria

UNIVERSIDAD
DE GRANADA

Nombre del equipo: AythaE

Ranking: 452 Puntuación: 0.8134

Sistemas Inteligentes para la Gestión en la Empresa

Máster en Ingeniería Informática

Curso 2016/17

Universidad de Granada

Nombre: Aythami Estévez Olivas
Email: aythae@correo.ugr.es

Índice

1. Exploración de datos

2. Preprocesamiento de datos

3. Técnicas de clasificación

4. Presentación y discusión de resultados

5. Conclusiones y trabajos futuros

Xgboost pero tal cual parece dar problemas, probar tecnicas que le vayan bien como

6. Listado de soluciones

La siguiente tabla recoge las distintas soluciones presentadas en Kaggle, tengo que mencionar inicialmente que son 11 filas en lugar de 12 a pesar de ser estos mis intentos en Kaggle. Esto se debe a que he subido la solución 3 dos veces debido a que se produjo un error durante la subida y lo volví a subir, por esto no la menciono en la tabla. Respecto a las posiciones del ranking son algo aproximadas ya que seleccionando una entrega como solución final no varia el ranking de Kaggle, por lo que he aproximado a las posiciones ocupadas por puntuaciones idénticas. Como software utilizado para todos los intentos se ha utilizado RStudio y los paquetes y funciones indicadas en la lista de abreviaturas.
La siguiente lista de abreviaturas por orden alfabético recoge los preprocesamientos y algoritmos utilizados para las distintas soluciones:

- AD1: Árbol de decisión usando el paquete y función `rpart` con el método "class" prediciendo la variable Survived usando Pclass, Sex, Age, SibSp, Parch, Fare y Embarked.
- AD2: Árbol de decisión usando el paquete y función `rpart` con el método "class" prediciendo la variable Survived usando Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, Title, FamilySize y FamilyID.
- CRF: Random Forest usando como unidad elemental conditional inference trees con la función `cforest` del paquete `party`, 2000 árboles y 3 variables aleatorias a elegir en cada nodo. Predice en función de Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, Title, FamilySize y FamilyID.
- CRFDS: Igual algoritmo que el previo pero creando dos modelos separados para hombres y mujeres, prediciendo hombres y mujeres por separado y luego uniendo los resultados.
- HM: Todos los hombres mueren.
- MSP3: Todas las mujeres se salvan menos las de la clase P3 que pagaron más de 20 por su billete.
- MSP3 + NHP3: Todas las mujeres se salvan menos las de la clase P3 que pagaron más de 20 por su billete y además todos los niños (menores de 18 años) hombres se salvan a expción de los de clase P3 que pagaron menos de 10 o más de 20.
- NA: Nada de preprocesamiento.
- RF: Random Forest del paquete homónimo con 2000 árboles prediciendo a partir de Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, Title, FamilySize y FamilyID2 (como Family ID pero considerando familia grande las de más de 3 miembros).
- TFSFID: Extracción del título social a partir del nombre, cálculo del tamaño de la familia en función de SibSp y Parch, además de generación de un ID de familias grandes (+ de 2 miembros).
- TFSFID + IEEF: Mismo preprocesamiento que el previo pero añadiendo imputación de la edad usando un árbol de regresión Anova a partir de Pclass, Sex, SibSp, Parch, Fare, Embarked, Title y FamilySize; imputación de los datos perdidos de embarque por el puerto más numeroso ("S") y de los datos perdidos del precio del pasaje por la mediana de su distribución.
- TFSFID2 + IEEF: Igual que lo anterior pero discretizando el tamaño de familia en "single" si < 2, "small" si > 1 y < 5 y "large" si > 4.
- TFSFID3 + IEEF: Igual preprocesamiento pero agrupando los títulos de manera distinta, considerando familia grande la que tiene 2 o más miembros y realizando la imputación de la edad usando el paquete `mice` y el método `rf`.

- TM: Como "algoritmo" se asume que todos mueren.

Nº de solución	Descripción Preprocesamiento	Algoritmos y Software	% Acierto en entrenamiento	% Acierto en test	Posición del Ranking
1	NA	TM	61,61616	62,679	13022 17/04
2	NA	HM	78,675	76,555	7742 18/04
3	NA	MSP3	80,8	77,99	4928 19/04
4	NA	MSP3 + NHP3	82,379	77,99	4928 19/04
5	NA	AD1	83,951	78,469	3525 21/04
6	TFSFID	AD2	85,522	79,426	1945 21/04
7	TFSFID + IEEF	RF	92,817	77,512	5997 22/04
8	**TFSFID + IEEF**	**CRF**	**85,634**	**81,34**	**452 22/04**
9	TFSFID2 + IEEF	CRF	85,634	80,383	819 22/04
10	TFSFID3 + IEEF	CRF	87,205	80,383	819 22/04
11	TFSFID + IEEF	CRFDS	85,185	81,34	452 22/04

Bibliografía

[1]: T. Stephens (n.d). Titanic: Getting Started with R. Recuperado el 25 de Abril de 2017, desde http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/

[2]: M.L. Risdal (2016). Exploring the Titanic Dataset. Recuperado el 25 de Abril de 2017, desde https://www.kaggle.io/svf/924638/c05c7b2409e224c760cdfb527a8dcfc4/__results__.html