Taller # 2 Introducción al Machine Learning para Economistas

Integrantes:

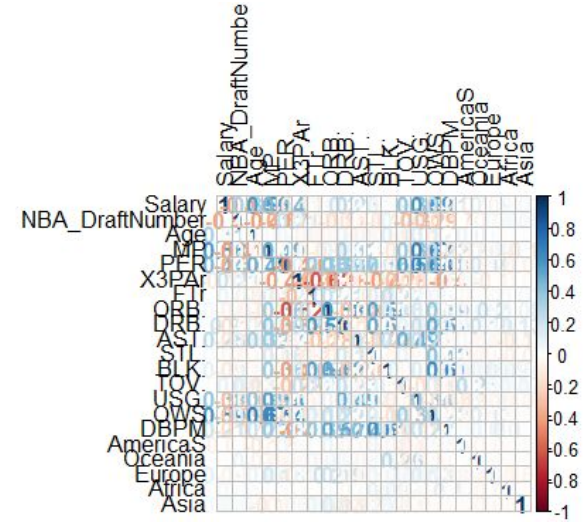
- José Sebastián Estepa Andrade
- Katherine Ramírez Cubillos

Procedimiento

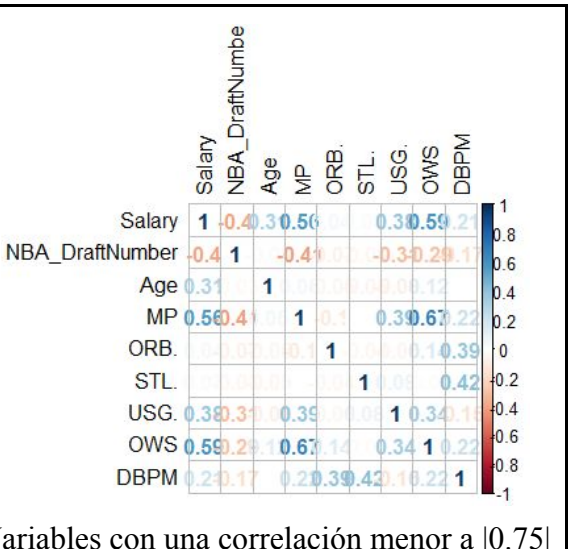
Al momento de cargar la base de datos, decidimos renombrar las filas con el nombre de cada jugador con el fin de que si algún jugador afectaba el modelo lo pudiésemos eliminar más facilmente. Adicionalmente, creamos 5 variables dummies de continente con los datos de lugar de nacimiento de los jugadores (dejando como caso base Norteamérica), limpiamos la base de datos de las variables omitidas y repetidas (Key Felder aparecía dos veces) y eliminamos las columnas Player, NBA_Country y Team, puesto que son columnas de texto.

Después de esto separamos los datos en entrenamiento y testeo por un factor del 75%.

Para evaluar la correlación entre las variables realizamos una prueba de no multicolinealidad, donde eliminamos variables hasta obtener aquellas con una correlación aceptable.



Todas las variables

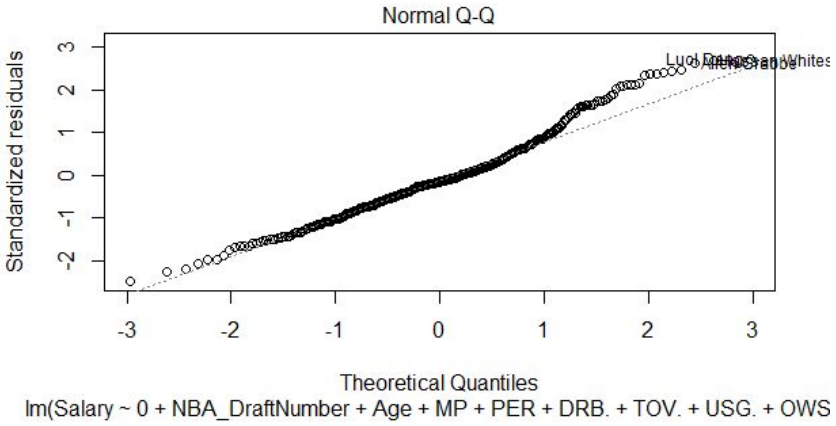


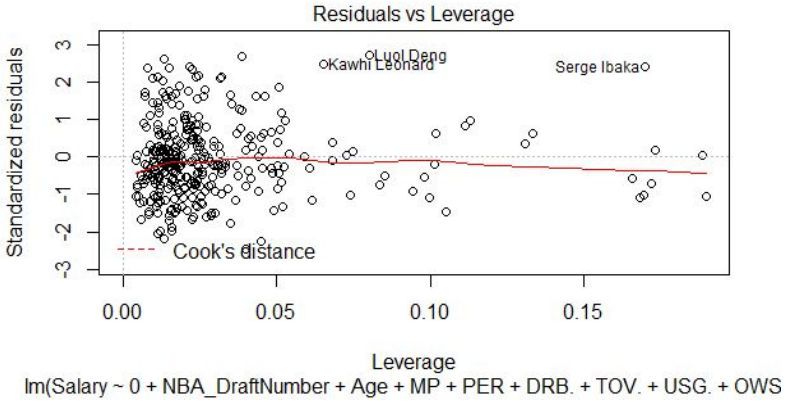
Variables con una correlación menor a |0.75|

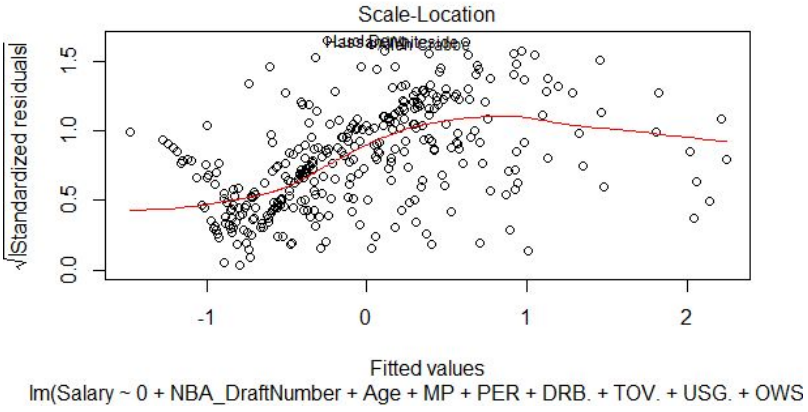
Regresión Lineal:

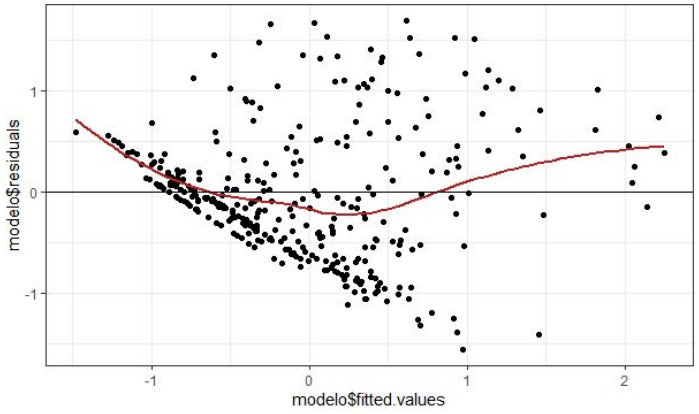
Se corrió el modelo con todas las variables y por medio del criterio AIC se seleccionaron las variables a trabajar. Una vez escogidas las variables se volvió a correr el modelo y se decidió eliminar el intercepto debido a la gran no significancia de este.

El modelo presentaba grandes problemas tanto de heterocedasticidad como de normalidad, y además de esto el error medio cuadrado de las predicciones era demasiado grande. Por esto se decidió normalizar la base de datos y volver a repetir los pasos ya hechos. Con esto se logró solucionar el problema del tamaño de los errores pero los problemas de normalidad y heterocedasticidad no desaparecieron. Como paso adicional se decidió eliminar datos atípicos de la base de datos por medio de un gráfico de influencias y eliminando 4-6 datos atípicos a la vez según la impresión de R. Se realizaron cinco tandas de eliminación de datos atípicos reduciendo la muestra de 480 datos a 455. Con esto se logró reducir bastante el problema de normalidad (aunque se logro eliminar). El problema de heterocedasticidad no fue posible siquiera reducirlo por lo que se decidió ignorar.









El modelo final fue:

lm(formula = Salary ~ 0 + NBA_DraftNumber + Age + MP + PER + DRB +
TOV + USG + OWS + DBPM + Africa, data = dt_train)

Con las variables de este modelo se procedió a correr los modelos de Máquina Soportada en Vectores, Árbol de Decisión y Random Forest.

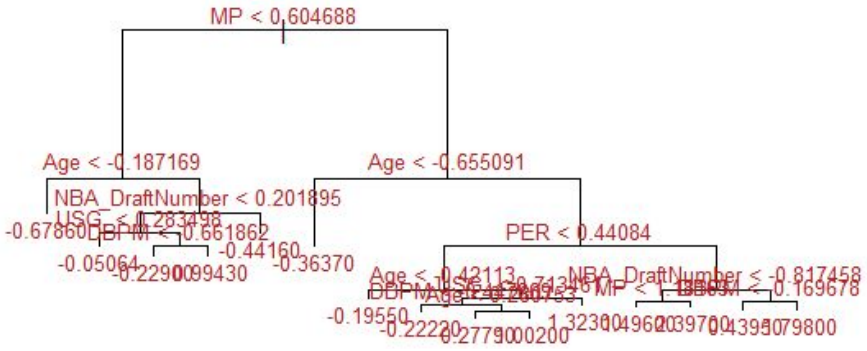
Técnica de regresión	**MSE**
Regresión Lineal	0.37
Árbol de decisión	0.47
Máquina soportada en vectores Number of Support Vectors	0.41 253
Bagging	0.35
Random Forest	0.36
Boosting cv_error n_arboles	**0.32** 0.3520579 337

Por lo que en este caso, el mejor modelo para predecir el salario (normalizado) es Boosting.

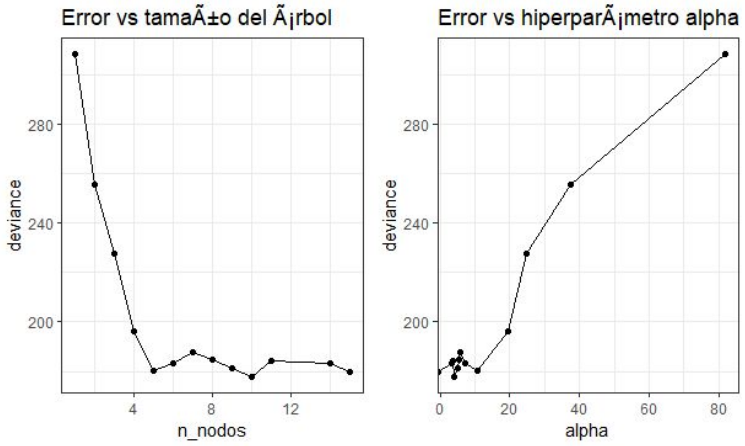
Resultados de las técnicas de regresión:

Árbol de decisión:

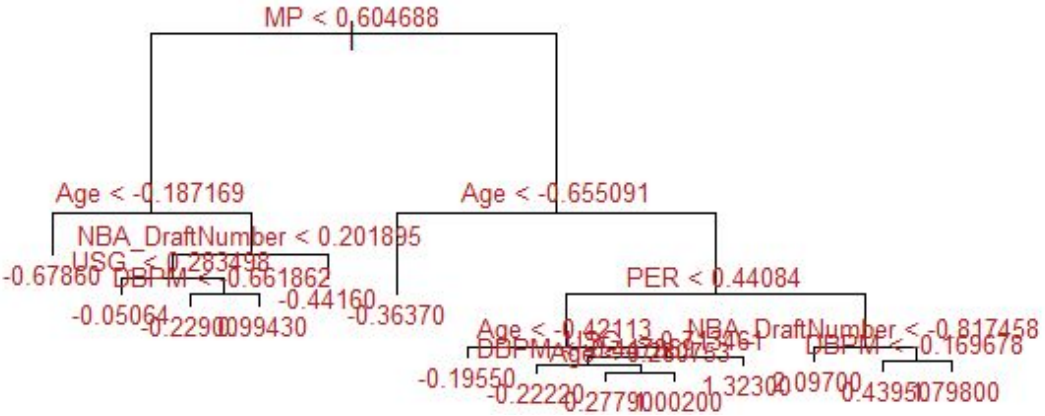
El primer árbol de decisión estimado fue el siguiente:



Para guiar el proceso de podado empleamos el método de clasificación de la tasa de error



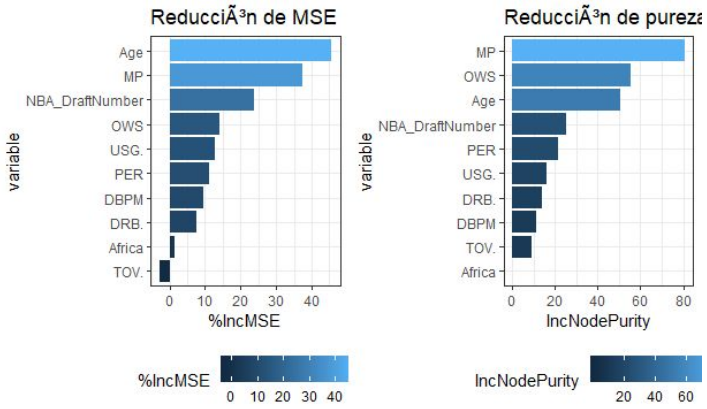
El árbol resultante del podado fue:



Bagging:

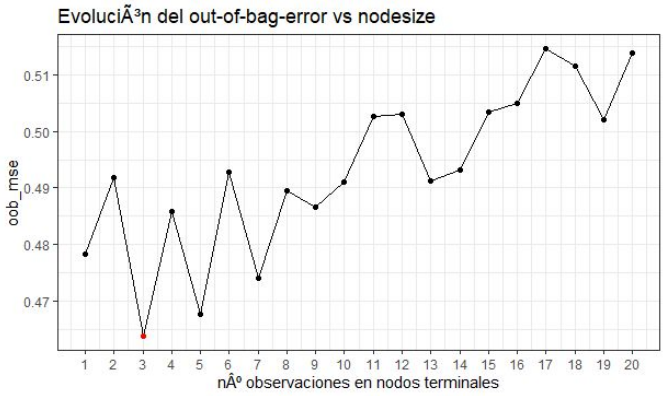
```
 randomForest(formula = Salary ~ 0 + NBA_DraftNumber + Age + MP +      PER + DRB. + TOV. + USG. +
OWS + DBPM + Africa, data = dt_train,      mtry = 13)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 0.3758362
                    % Var explained: 55.82
```

La salida muestra el número de árboles generados (500), el Mean of Squared Residuals y el porcentaje de varianza que el modelo es capaz de explicar. Por otra parte, en este modelo, en los árboles generados en el proceso de bagging las variables Age y MP son las que más influyen en el modelo.

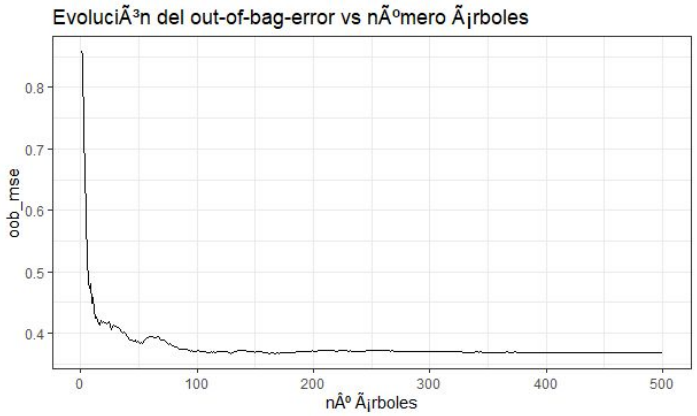


Random Forest:

Al momento de establecer este modelo utilizamos 500 árboles. Nuestro modelo identifica el 3 como el número óptimo de observaciones mínimas que deben contener los nodos terminales.

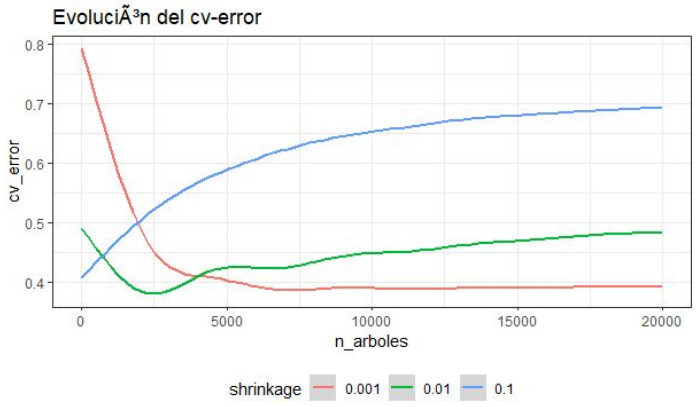


Por otra parte, para identificar el número óptimo de árboles graficamos la evolución del out-of-bag-error en función del número de árboles. En nuestro modelo, a partir de aproximadamente 100 árboles la precisión del modelo se estabiliza.

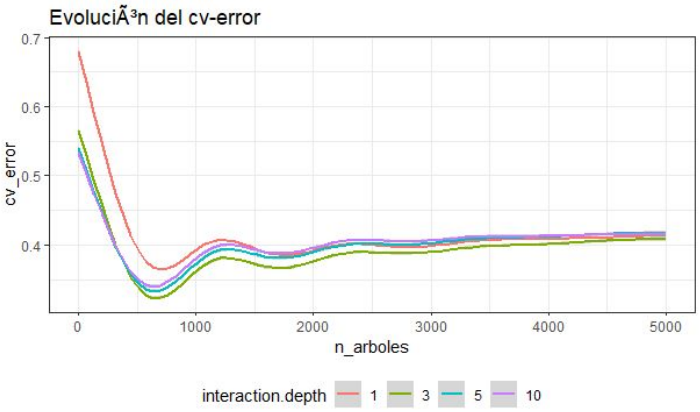


Boosting:

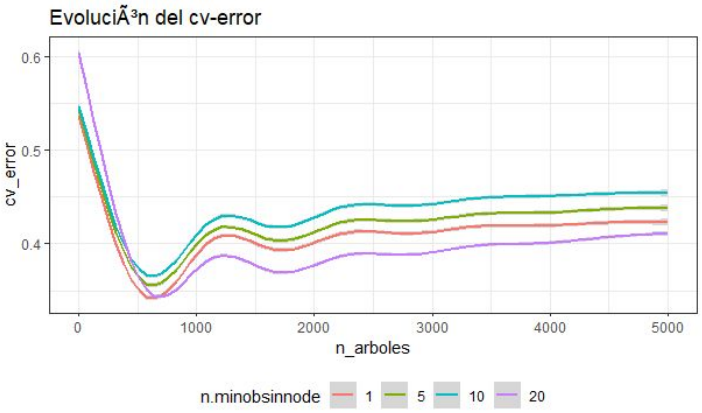
Los resultados de nuestro modelo boosting son los siguientes:



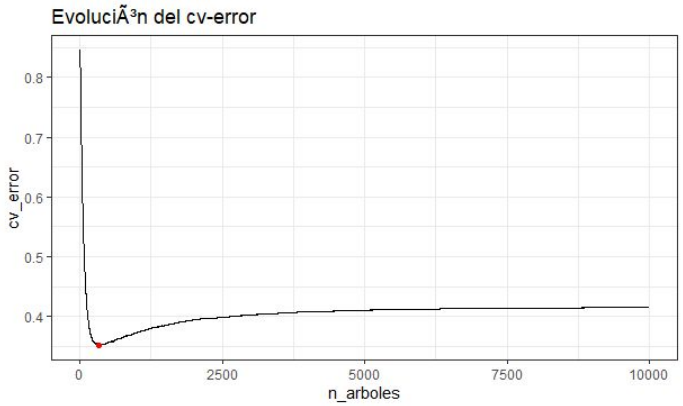
Complejidad del modelo:



Número mínimo de observaciones por nodo:



Número de árboles:



El modelo muestra que un sistema de 337 consigue el menor error de validación cruzada de 0.3520579.

En cuanto a los predictores, en nuestro modelo a medida que aumenta la variable MP aumenta el salario del jugador. Lo mismo ocurre con la variable edad.