Competencia Métodos

Javier Arturo Rozo Alzate - jarozoa@eafit.edu.co
Alejandro Palacio Vásquez - apalac19@eafit.edu.co
Liceth Cristina Mosquera Galvis - lcmosquerg@eafit.edu.co
Cristian David Muñoz Mora - cdmunozm@eafit.edu.co
Programa: Métodos Estadísticos Avanzados
Docente:
Andrés Ramirez Hassan - aramir21@eafit.edu.co
12 de octubre de 2019

***Resumen*—En este documento se presenta la metodología utilizada para abordar el problema de selección de variables y predicción planteado en la materia, para 3 bases de datos diferentes. Se inicia con un análisis exploratorio de la información, para posteriormente realizar con diferentes modelos estadísticos la selección de las variables explicativas más significativas y finalmente proceder a evaluar los resultados de predicción.**

Palabras Clave: Selección de variables.

I. Introducción

En estadística, los métodos de regularización son utilizados para la selección del modelo y para evitar el sobreajuste en las técnicas predictivas. Por ende, al abordar el trabajo, se quiere ofrecer una revisión general de la metodología y diferentes fases del proceso de selección de variables de una base de datos con alta dimensionalidad, así como de los criterios de selección y descripción de las diferentes técnicas que pueden utilizarse en la investigación de carácter aplicativo a la ciencias de los datos con diferentes metodologías y con el uso apropiado de las técnicas estadísticas, que ha de ser acorde con el tipo de información disponible.

Se usan diferentes metodologías robustas tanto bayesianas como frecuentistas para la correcta selección de variables, ya que es necesario seleccionar las mejores variables predictivas o registros auxiliares, también llamados regresores. Y, de esta forma, seleccionar la mejor ecuación de regresión de entre todas las posibles combinaciones. Donde al final, lo que se quiere es crear el modelo más simple e interpretable posible.

Uno de los objetivos es obtener un modelo parsimonioso, es decir, ajustar bien los datos a la variable de respuesta, pero usando la menor cantidad posible de variables explicativas o de regresores. Donde, la selección de variables y multicolinealidad son dos problemas que se pueden tratar de manera simultánea. Bajo este escenario se ubica el objetivo central de este trabajo.

II. Descripción del Problema

El objetivo de la investigación es estudiar y determinar las influencias significativas que ejercen las distintas variables, en correspondencia con el rendimiento final, determinando esencialmente la cantidad de variables que logren, de manera óptima, predecir el comportamiento de la variable respuesta. Donde, lo que se quiere al final, es analizar la capacidad predictiva general de los tres modelos con respecto a cada una de las bases de datos. En la base de datos continua y conteo, se medirá de acuerdo con el error cuadrático medio y para la base de datos binaria, se medirá de acuerdo de acuerdo con el Accuracy de los resultados (verdaderos positivos más verdaderos negativos dividido el tamaño de la muestra). La capacidad predictiva específica para la base de datos Continua será la correcta clasificación de los valores inferiores y superiores a -1, para la base de datos Binaria el área bajo la curva ROC y para la base de datos de Conteo, será la correcta clasificación de los valores iguales a 0 y mayores a 0. Para la correcta selección de variables, se medirá de acuerdo con el modelo ya establecido por el profesor y que tan similar es el modelo seleccionado con el establecido.

III. Análisis Exploratorio de los Datos

Inicialmente se inicia explorando cada una de las tres bases de datos con las que se cuenta para la realización del trabajo, donde la diferencia principal entre estas bases de datos es la variable de respuesta dependiente y. La primera base de datos tiene como variable dependiente una variable continua, la segunda tiene un comportamiento binario (0-1) y la última tiene como variable dependiente un comportamiento tipo conteo (0-7).

- La variable dependiente Continua de la primera base de datos tiene un comportamiento aparente de distribución normal como lo muestra la gráfica de la distribución de los datos. Donde, el número menor de la distribución es -6.4341 y el número máximo es 6,8003.

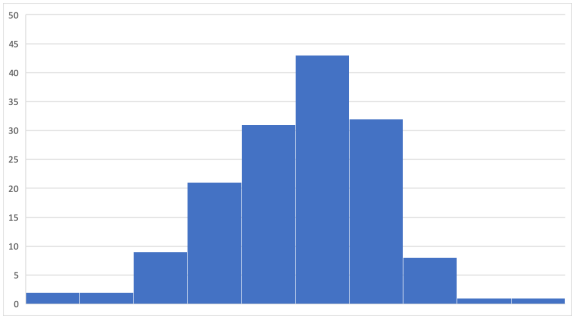

Figura 1: Distribución de la variable dependiente continua

- La variable dependiente binaria de la segunda base de datos, tiene un comportamiento desbalanceado em sus datos ya que cuenta con menos de la mitad de ceros que de unos.

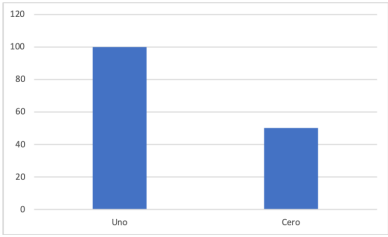

Figura 2: Cantidad de datos en la variable dependiente binaria

- La variable dependiente de conteo de la tercera base de datos, que va desde cero hasta siete, así como la base de datos anterior, es un conjunto de datos desbalanceado, donde el número 6 y 7 no representan ni el 1 % del total de los datos, convirtiéndose en datos atípicos dentro del conjunto de datos. Estos datos atípicos podrían ser complejos a la hora de abordar el algoritmo de selección de variables ya que, algunos algoritmos son pobres en recursos para manejar este tipo de insuficiencia de datos. Por ende, el algoritmo de clasificación podría clasificar de manera errónea estos casos.

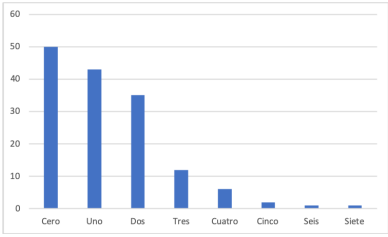

Figura 3: Cantidad de datos en la variable dependiente conteo

Las bases están conformadas por 150 observaciones y 34 variables, de las cuales la primera columna de las bases es el id que funciona como fila identificadora de cada una de las observaciones dentro de la base. La segunda columna es la variable dependiente (como explicamos anteriormente) en cada una de las bases y el resto de las 32 columnas son las variables independientes de las bases de datos. Al final, las tres bases de datos son iguales en sus 32 variables independientes y la única diferencia es la variable dependiente de las bases de datos.

Las 32 variables independientes se comportan de la siguiente forma: 13 de ellas tomas valores continuos, 4 valores discretos o de conteo y otras 15 son variables logísticas o binarias 0-1. En el siguiente gráfico, se presenta el comportamiento de las variables continuas y discretas en función de su correlación para detectar multicolinealidad.

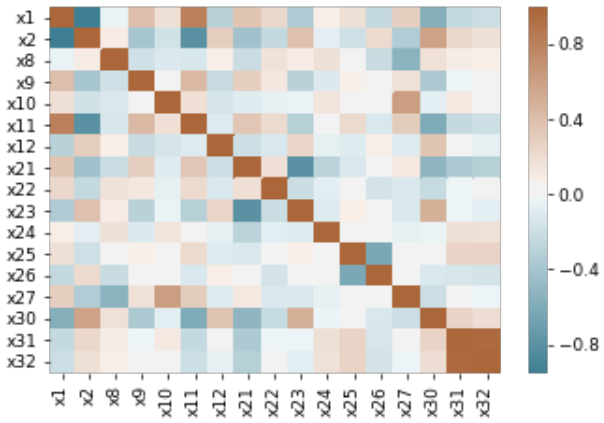

Figura 4: Correlación datos continuas y discretas

De lo anterior, se puede observar que existe alta correlación entre las variables x_{29}, x_{30} x_{31} y x_{32} y esta es información relevante para los análisis que realizaremos posteriormente.

A continuación, se presenta la distribución de estas variables.

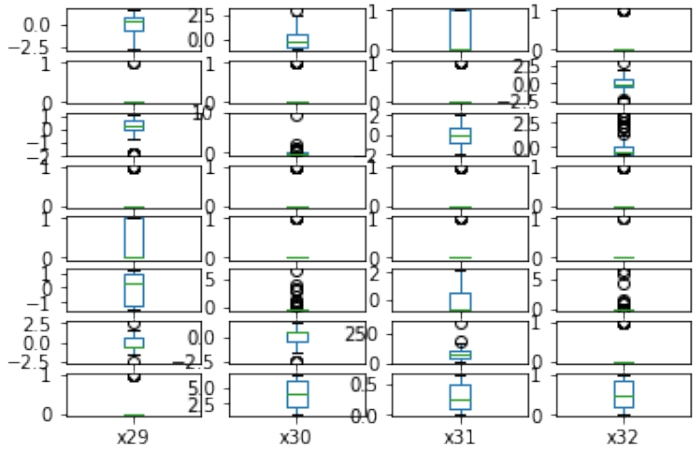

Figura 5: Correlación datos continuos

IV. Modelos Evaluados para la Selección de Variables

IV-A. Lasso (Least Absolute Shrinkage and Selection Operator)

Como método de regresión regularizada, este modelo genera un análisis de regresión que realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este [1].

Donde, en la fórmula, se tiene integrada la restricción que tiene la función para la penalización de los coeficientes de la función por medio del parámetro λ. Para lograr de esta forma estabilizar las estimaciones y predicciones y, por ende, realizar la selección de variables.

Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y al mismo tiempo produce

modelos más interpretables y simples por la reducción de algunos coeficientes a cero.

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{K} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{K} |\beta_j| \right\} \quad (1)$$

Utilizando Lasso, es posible obtener un modelo con buena precisión y que sea interpretable, pero este método también tiene varias limitaciones como lo son las siguientes:

- En el caso $P > n$ Lasso selecciona a lo sumo n variables antes de saturarse, debido a la naturaleza del problema de optimización convexa y esto podría ser una limitación para un método de selección de variables. Además, Lasso no está bien definido a menos que el límite de la norma L1 de los coeficientes sea menor que un cierto valor.
- Si existe un grupo de variables entre las cuales las correlaciones por parejas son muy altas, entonces Lasso tiende a seleccionar sólo una variable del grupo, sin importarle cuál de ellas selecciona.
- Para el caso $n > P$, si existe una alta correlación entre los predictores, se ha observado que, en general, la predicción a través de regresión Ridge resulta más óptima que la obtenida a través de Lasso.

IV-B. Lars

"Least Angle Regression" (Efron et al., 2004) es un modelo que puede verse como un especie de forward stepwise regression, ya que usa una estrategia similar de ir agregando variables al modelo, basado en la correlación con los residuales.

Estructura del algoritmo

1. Estandarizar los predictores para que tengan media cero y norma unitaria, se calculan los residuales.
2. Se encuentra el pedictor X_j que tenga mas correlación con r
3. Se mueve B_j desde un valor de 0 hasta su coeficiente en mínimos cuadrados, $< X_j, r >$ Hasta que otro competidor X_k tenga tanta correlación con el residual como tiene X_j.
4. Se mueve B_j y B_k en la dirección definida por el coeficiente conjunto de mínimos cuadrados, del residual, hasta que otro competidor X_l tenga tanta correlación como la actual sobre los residuales.
5. se continua hasta que todos los predictores estén en el modelo

IV-C. Elastic Net

Hui Zou and Trevor Hastie [2] proponen en 2005 una nueva técnica de regularización y selección de variables conocido como Elastic Net, la cual retiene las ventajas de Lasso, hace automáticamente selección de variables y contracción continua, y al mismo tiempo supera algunas de sus limitaciones. Con este nuevo método se puede seleccionar grupos de variables correlacionadas. Este método es particularmente útil cuando el número de predictores (P) es mucho más grande que el número de las observaciones (n). En primer lugar, los autores define Naive Elastic Net(red elástica simple) que es un método de mínimos cuadrados penalizado utilizando una penalización nueva de Elastic Net.

Donde, su función de penalización es la siguiente:

$$\lambda \sum_{j=1}^{K} (\alpha|\beta_j| + (1-\alpha)\beta_j^2). \quad (2)$$

Siendo esto, una combinación entre los modelos de regresión entre lasso y ridge. Donde al final lo que hace la regresión es la selección de variables como el Lasso y reduce los coeficientes de los predictores correlacionados como lo hace Ridge. De igual forma, tiene considerables ventajas computacionales sobre las penalizaciones Lq.

IV-D. BMA

Bayesian Model Averaging (BMA) es una aplicación de inferencia bayesiana para los problemas de selección de modelos que se enfoca en explicar la incertidumbre del modelo [3]. En particular, se considera la problemática de selección de variables, dada la incertidumbre de los regresores, en un marco de regresión donde existen K variables explicativas, lo que implica $M = M_1, M_2, ..., M_{2^k}$ posibles modelos indexados por los parámetros θ_m, $m = 1, 2, ..., 2^k$.

Generalmente, en la práctica estadística, se suele ignorar la incertidumbre del modelo y los analistas de datos seleccionan un modelo, perteneciente a alguna clase de modelos, para proceder con el supuesto de que el modelo seleccionado es el generador de los datos. Lo anterior ignora la incertidumbre en la selección del modelo, lo que lleva a inferencias estadísticas o construcción de modelos que pueden estar muy confiadas pero que presentan errores en la calibración de este.

BMA proporciona un mecanismo coherente para tener en cuenta esta incertidumbre del modelo al derivar las estimaciones de los parámetros, marginando a los modelos para derivar densidades posteriores en los parámetros del modelo que explican la incertidumbre del modelo, de la siguiente manera:

$$\pi(\theta|y) = \sum_{m_i} \pi(m_i|y)\pi(\theta|y, m_i) \quad (3)$$

Donde, m_i es el conjunto de posibles modelos, $\pi(m_i|y)$ es la posterior probability del modelo m_i y $\pi(\theta|y, m_i)$ es la posterior density de los parámetros condicionados al modelo m_i. $\pi(\theta|y, m_i)$ es un proxy adecuado para la información sobre los parámetros θ cuando $\pi(\theta|y, m_i) \approx 1$.

Un estadístico importante que se obtiene en el proceso del BMA es la posterior inclusion probability (PIP) asociada a cada variable x_l, $l = 1, 2, ..., k$ y se define como:

$$PIP(x_i) = \sum_{m_i} \pi(m_i|y) * I_{l,m} \quad (4)$$

Donde, $I_{l,m} = 1$ si $x_l \in m_i$ y 0 en otro caso.

Revisando la literatura [4], respecto a las PIP de las x_l se sugiere lo siguiente:

- PIP<0.5, es evidencia en contra del regresor.
- 0.5$\leq PIP <$0.75, es evidencia débil a favor del regresor.
- 0.75$\leq PIP <$0.95, es evidencia positiva a favor del regresor.
- 0.95$\leq PIP <$0.99, es evidencia fuerte a favor del regresor.
- PIP$\geq$ 0,99, es evidencia muy fuerte a favor del regresor.

BMA permite incorporar la incertidumbre del modelo en un marco de regresión y cuando se desee seleccionar sólo un modelo existen dos alternativas: el modelo con la mayor posterior model probability y el modelo de probabilidad mediana, siendo este último modelo aquel que incluye cada predictor x_l que tenga una PIP superior a 0.5

IV-E. XGboost

Este modelo está basado en árboles de decisión unido a un algoritmo de aprendizaje automático que usa un marco de potenciación del gradiente. Se construyen varios arboles en espacios aleatorios del espacio de variables, árboles en distintos subespacios generalizan su clasificación de manera complementaria [5] y [6].

El boosting es una técnica secuencial que funciona según el principio del conjunto (ensemble). Combina un conjunto de aprendices (learners) débiles y ofrece una precisión de predicción (accuracy) mejorada. En cualquier instante t, los resultados del modelo se pesan en función de los resultados del instante t-1 anterior. Los resultados pronosticados correctamente reciben un peso menor y los clasificados erróneamente tienen mayor peso. Esta técnica se sigue para un problema de clasificación, mientras que se usa una técnica similar para la regresión [7]

Dentro de las ventajas de este modelo está que su regularización ayuda a reducir el sobre ajuste, a pesar que es proceso secuencial se puede hacer procesamiento en paralelo, tiene alta flexibilidad para definir los objetivos de optimización y los criterios de evaluación, tiene una rutina incorporada para manejar los valores perdidos, la poda de árboles la hace eliminando divisiones donde no haya ganancia, además que permite ejecutar una validación cruzada en cada iteración.

Otras de las ventajas es que no requiere escalar las variables, puede tratar con entradas irrelevantes, es interpretable si es pequeño, puede manejar predictores tanto cuantitativas como cualitativas, sin embargo no se puede extraer combinaciones lineales de las variables y tiene un bajo poder de predicción cuando hay alta varianza.

IV-F. RFE

Recursive feature elimination. (Eliminación de características recursivas). Detrás de esta eliminación, creada por la librería Scikit Learn, la cual es una librería multicolaborativa. Donde es abierta a mejoras por la comunidad colaborativa de desarrolladores. Esta metodología aunque esta basada en la eliminación de variables, presenta dificultades cuando se analizan problemas con variables altamente redundantes. Donde originalmente esta metodología esta basada en la caracteristicas menos importantes y de esa forma ir eliminando variables hasta quedar con el pequeño y mejor modelo con las variables más aportantes.

V. Base de datos Binaria

V-A. Modelos para la Selección de Variables

Una vez se realizó el preprocesamiento de la base de datos, en el que se estandarizó las variables continuas y de conteo, se procede a hacer la selección de variables. Para este fin se evaluaron diferentes modelos tanto frecuentistas como bayesianos, obteniendo los siguientes resultados:

Mejor Modelo Frecuentista - GLM con Regularización con Lasso.

Para este modelo se utilizo la librería Glm net, desarrollada por Hastie y su equipo, para iniciar el proceso se identificó a través de cross validation, el valor de λ que dentro del proceso minimiza los errores.

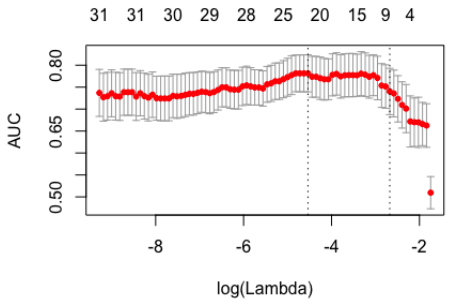

Figura 6: lambda por C.V. minimizando AUC

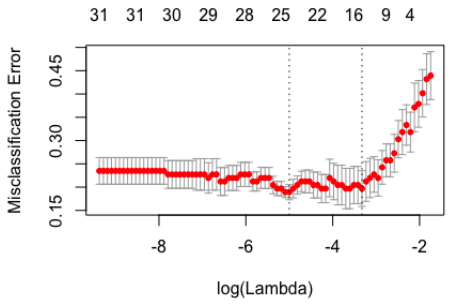

Figura 7: lambda por C.V. minimizando error de clasificación

Resultados de Lambda por validación cruzada

	Misclasification error	AUC
Lambda min	0.006	0.1
Lambda 1se	0.35	0.6

Figura 8: valor de lambda hallados

Una vez se definen los posibles valores de λ, se realiza el modelo calculando los cuatro modelos y se evalúan contra el conjunto de datos de prueba, manteniendo el modelo con las variables que presentan los mejores indicadores de predicción.

Modelo propuesto bajo Lasso:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \hat{\beta}_3 X4_i + \hat{\beta}_4 X17_i + \hat{\beta}_5 X19_i + \hat{\beta}_6 X23_i + \hat{\epsilon}_i \quad (5)$$

Mejor modelo Bayesiano - Bayesian Model Averaging (BMA).

Esta técnica calcula X^K modelos, donde K es el número de variables, calculando un promedio ponderado con pesos basados en la posterior model probabilities.

Para la selección de variables se eligen las que se se incluyen el mayor número de veces en los mejores modelos.

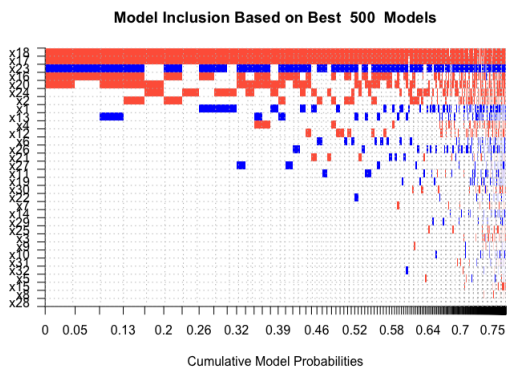

Figura 9: valor de lambda hallados

Dado que no tenemos conocimiento previo de los modelos, asignamos una probabilidad uniforme para las prior.

Resultados PIP

	PIP	Post Mean	Post SD	Cond.Pos.Sign	Idx
x18	0.90433333	-3.541975e-01	1.793943e-01	0.00000000	18
x17	0.82400000	-3.438396e-01	2.161594e-01	0.00000000	17
x23	0.76233333	2.796787e-01	1.980217e-01	1.00000000	23
x2	0.47200000	-1.396763e-03	1.852611e-03	0.00000000	2
x20	0.44666667	-1.803808e-01	2.441721e-01	0.00000000	20
x24	0.37666667	-2.376617e-01	3.587153e-01	0.00000000	24
x16	0.29566667	-8.520496e-02	1.512753e-01	0.00676437	16
x4	0.24866667	-5.451218e-02	1.139175e-01	0.00000000	4
x1	0.24633333	2.758645e-03	6.304182e-03	0.98105548	1

Figura 10: Información sobre coeficientes.

Bajo la premisa de tomar como evidencia a favor del regresor un PIP mayor al 0.4, las variables seleccionadas son la x_2, x_{11}, x_{18} y x_{23}.

Modelo propuesto bajo BMA:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X2_i + \hat{\beta}_2 X17_i + \hat{\beta}_3 X18_i + \hat{\beta}_4 X20_i + \hat{\beta}_5 X23_i + \hat{\epsilon}_i \quad (6)$$

V-B. metodología seleccionada y resultados

Las variables sobre las que se observan los mejores resultados son las obtenidas con el modelo BMA, siendo las variables seleccionadas la x_2, x_{11},x_{17}, x_{18}, x_{20} y x_{23}.

Para realizar la predicción de los datos, se usa un modelo logit obteniendo los siguientes resultados:

Split	Accuracy	AUC
1	77%	87%
2	77%	91%
3	82%	88%
4	71%	77%
5	80%	83%
6	80%	86%
7	73%	82%
8	71%	79%
9	71%	80%
10	80%	79%
Promedio	76%	83%

Figura 11: resultados Accuracy y AUC bajo diferentes particiones entrenamiento, prueba

Análisis de multicolinealidad por vif

x18	x17	x23	x20	x2
1.189004	2.047912	1.329384	1.099455	1.963807

Figura 12: VIF Multicolinealidad modelo seleccionado

Al evaluar si las variables seleccionadas presentan multicolinealidad, se evidencia que no es así, ya que se observa su variance inflation factor (VIF) inferior al 2.5, por lo que podemos afirmar que no hay evidencia estadística de presencia de multicolinealidad.

Distribución de los errores.

VI. Base de datos Conteo

VI-A. Modelos para la selección de variables

Para la selección de variables de la base de datos de conteo se utilizó el modelo XGBoostRegresión con la función objetivo de minimizar el error cuadrático medio (Mean Squared Error).

Se probaron varios hiperparámetros para encontrar cuales eran los que mejor podían ayudar para hallar el mejor resultado de regresión del conjunto de datos.

Los hiper parámetros que se cambiaron fueron:

- colsample bytree: ente 0.4, 0.6 y 0.8, este indica la porción de columnas que tomará como muestra aleatoria para cada árbol.

- gamma: Se entrenó con valores 0, 0.03, 0.1 y 0.3 el cual especifica la perdida mínima requerida para hacer un división (split).
- min child weight:Se entrenó con valores 1.5, 6 y 10, define la suma mínima de pesos de todas las observaciones requeridas en un hijo. Ayuda a prevenir la sobre estimación.
- learning rate: Se entrenó con valores 0.1 y 0.07, tasa de aprendizaje que hace el modelo más robusto al reducirlos pesos de cada paso.
- max depth':máxima profundidad del árbol, se utiliza para controlar el sobre entrenamiento, una mayor profundidad permite aprender relaciones muy específicas para una muestra particular en este caso se entrenó con 3 y 5.
- n estimators':número de árboles a ajustar secuencialmente 10000.
- reg alpha': regularización L1 para reducir el sobre ajuste análogo a regresión Lasso, controla la complejidad, se entrenó con 1e-5, 1e-2 y 0.75. Puede ser usado en caso de alta dimensión.
- reg lambda': Termino de regularización L2 (análogo a regresión Ridge, también usado para reducir el sobre entrenamiento, se entrenó con 1e-5, 1e-2 y 0.45.
- subsample':porción de observaciones para obtener una muestra aleatoria para cada árbol en este caso se usaron 0.6 y 0.95.

Después del entrenamiento las mejores se escogen los mejores parámetros con los que se evalúa la importancia de las variables. J.Brownlee2019

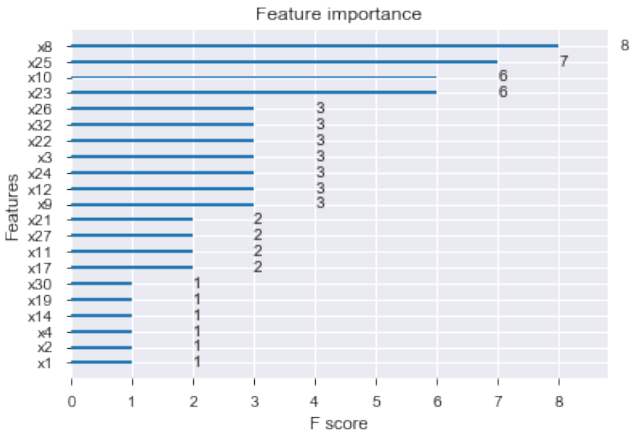

Figura 13: Importancia de las variables

VI-B. *Variables y Metodología Seleccionada*

Con las veintiún (21) variables resultantes después del entrenamiento se prueban verificando cuáles mejoran el error cuadrático medio y por principio de parsimonia también se eliminan aquellas que no cambian el resultado.

Se utilizó el metodo XGBoost Regressor con objetivo 'count:poisson' dado que la variable dependiente es de conteo con el 85 % de los datos entre las 0, 1 y 2.

Las variables seleccionadas fueron x_3, x_8, x_{23} y x_{25}. Donde, x_8 y x_{23} son variables continuas, x_3 es una variable logística y x_{25} es una variable de conteo.

Con estas variables, se obtiene un MSE es de 1.66

VI-C. *Análisis de los resultados*

Estas variables se escogieron teniendo en cuenta los objetivos del menor error cuadrático medio y la correcta predicción ente los valores mayores a cero y los valores iguales a cero.

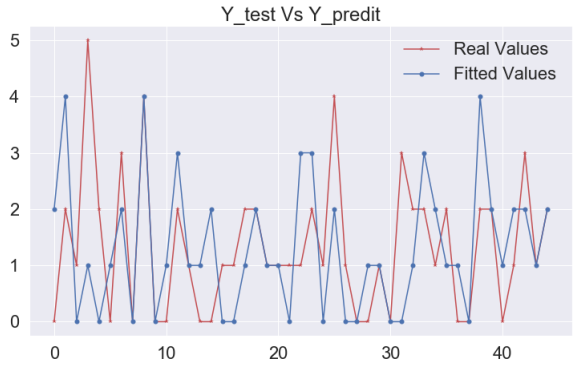

Figura 14: Resultado del test de conteo

VII. Base de datos Continua

VII-A. *Modelos para la Selección de Variables*

En el caso de la variable continua, se trabajaron las siguientes opciones de modelos para la selección de variables:

- Lasso
- Elastic Net
- Ridge
- BMA

Trabajando con todos los datos, en primer lugar y al ejecutar el modelo Lasso para selección de variables se obtienen los siguientes resultados:

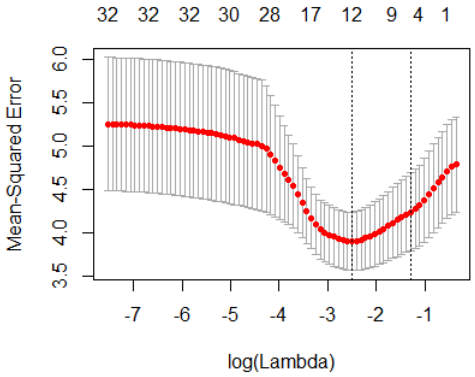

Figura 15: Resultados Lasso Variable Continua

Donde se obtiene que el λ que minimiza el MSE es de 0.082 y que las variables seleccionadas son $x_1, x_{10}, x_{23}, x_{24}, x_{27}, x_{31}, x_{32}, x_6, x_{13}, x_{20}, x_{25}$ y x_{29}.

Por otro lado, al correr el modelo Elastic Net con un α de 0.3 se obtienen los siguientes resultados:

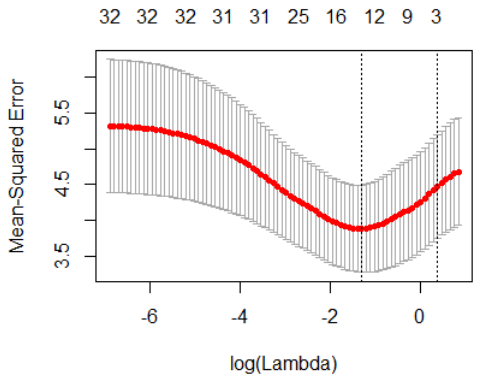

Figura 16: Resultados Elastic Net Variable Continua

Donde se obtiene que el λ que minimiza el MSE es de 0.273 y que las variables seleccionadas son $x_1, x_{10}, x_{23}, x_{24}, x_{27}, x_{31}, x_{32}, x_3, x_6, x_{13}, x_{20}, x_{25}$ y x_{29}.

Los resultados obtenidos entre estos dos modelos iniciales son muy similares en cuanto a la selección de variables, sólo que el Elastic Net propone una variable más que es x_3 sin embargo, y como se presentó en secciones anteriores, existe cierta correlación entre algunas de las variables seleccionadas por lo que se decide utilizar el modelo Ridge dado que esta regresión reduce en mayor proporción las direcciones de las variables en el espacio X que tienen variaciones más pequeñas. En ese sentido, al tener dos variables correlacionadas dentro del modelo, aquella cuyo β en la regresión Ridge sea mayor en magnitud absoluta tiene mayor fuerza explicativa en el modelo con menor variación.

Al correr la regresión Ridge con las variables seleccionadas por el modelo Elastic Net se obtiene:

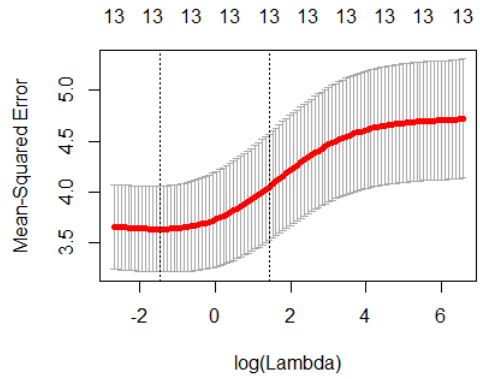

Figura 17: Resultados Ridge Variable Continua

Donde se obtiene que el λ que minimiza el MSE es de 0.234 y que los β del modelo son:

```
x1   0.1804685
x10  0.3479925
x23  0.5744660
x24 -0.1585666
x27  0.2116843
x31 -0.2598044
x32 -0.1755925
x3  -0.1818889
x6   0.7656372
x13 -1.1439327
x20  2.0720748
x25  0.4394850
x29  0.4840170
```

Figura 18: Resultados Ridge Variable Continua

Donde podemos observar que entre las variables correlacionadas x_{31} y x_{31}, es x_{31} quien obtiene el mayor β en magnitud y por ende permanece en las variables seleccionadas.

Por último, corrimos la metodología BMA con el fin de obtener el modelo de probabilidad mediana, siendo este modelo aquel que incluye cada predictor x_l que tenga una PIP superior a 0.5, demostrando así evidencia a favor de dicha variable para pertenecer al mejor modelo. Las variables obtenidas bajo este enfoque fueron: $x_{10}, x_{13}, x_{20}, x_{23}, x_{25}$ y x_{31}.

VII-B. Variables y Metodología Seleccionada

En resumen, se obtuvieron los siguientes resultados para la selección de variables:

Modelo	Variables
Lasso	$x_1, x_{10}, x_{23}, x_{24}, x_{27}, x_{31}, x_6, x_{13}, x_{20}, x_{25}$ y x_{29}
Elastic Net	$x_1, x_{10}, x_{23}, x_{24}, x_{27}, x_{31}, x_3, x_6, x_{13}, x_{20}, x_{25}$ y x_{29}
BMA	$x_{10}, x_{13}, x_{20}, x_{23}, x_{25}$ y x_{31}

Al final y para realizar las validaciones seleccionamos las variables que se obtuvieron con la metodología BMA.

VII-C. Análisis de los Resultados

Corrió el modelo GLM con la variable y como dependiente en función de todas las variables que seleccionamos con todos los datos, es decir, las 150 observaciones que teníamos disponibles. Así, se obtuvieron los siguientes resultados:

Partición	MSE	Accutacy
1	3.56	0.75
2	3.29	0.77
3	3.47	0.82
4	3.83	0.66
5	3.29	0.68
6	3.57	0.7
7	2.43	0.77
8	3.66	0.77
9	4.31	0.66
10	4.11	0.68

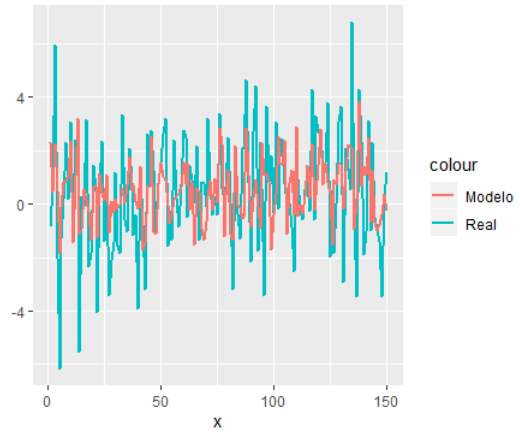

Figura 19: Resultados GLM Variables Seleccionadas

Con lo anterior, para la capacidad predictiva general, se obtuvo un MSE de 3.26 mientras que, para la capacidad predictiva específica, en términos de la correcta clasificación de los valores inferiores y superiores a -1, se obtiene un Accuracy de 0.75

Este último ejercicio se realizó de manera iterativa, generando particiones sobre los datos para obtener un conjunto de datos de entrenamiento (con el 70 % de los datos) y uno de validación con el 30 % restante. Esta partición la realizamos 10 veces para analizar el comportamiento del MSE y del Accuracy y el resultado se presenta en la tabla VII-C.

En general, observamos que los resultados oscilan entre unas bandas acotadas y que la volatilidad se explica principalmente por la partición en los datos que estamos realizando. Donde si, trabajáramos con una partición de 80-20 los resultados serían más estables. Sin embargo, nos sentimos tranquilos con los resultados que estas variables arrojan.

VIII. Conclusiones

Como conclusión se presentó en este trabajo la manera de abordar el problema de selección de variables en tres casos de variables dependientes de diferente tipo: continua, binaria y conteo, en los que no se conocían a priori los procesos generadores de los datos.

Se aplicaron técnicas como Lasso, Ridge, Elastic Net, XgBoost y BMA con dicho fin y se analizaron los resultados que en general, fueron acertados dentro de los parámetros establecidos para la medidción de la capacidad predictiva general y específica de los modelos.

Referencias

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of Royal Statistical Society: Series B (methodology)*, vol. 67, no. 1, pp. 91–108., 1996.

[2] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. R. Statist. Soc. B*, vol. 67, no. 2, p. 301–320, 2005.

[3] T. M. Fragoso and F. L. Neto, “Bayesian model averaging: A systematic review and conceptual classification,” *Statistical Science*, 2015.

[4] B. M.D and B. J.O, “Optimal predictive model selection1,” *The Annals of Statistics*, vol. 32, no. 3, pp. 870 – 897, 2004.

[5] J. Brownlee, *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016.

[6] J. Brownlee, “Feature importance and feature selection with xgboost in python,” 2019.

[7] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,”