
quantil

Predicción del desempeño académico en la Universidad de los Andes

Hamadys L. Benavides Gutiérrez
Quantil-Universidad de los Andes

15 de febrero de 2018

Generalidades

- Convocatoria Icfes.
- Uso de la base de datos Icfes y Universidad de los Andes.
- Pregunta de investigación:
  - ¿Cuáles son los factores que más contribuyen a predecir el éxito académico de los estudiantes en los cursos básicos de la Universidad de los Andes en el período 2015-2016?
- Objetivo general:
  - Identificar las variables relevantes en la predicción del desempeño académico de los estudiantes admitidos en la Universidad de los Andes entre el período 2015-2016, usando técnicas de aprendizaje de máquinas.

Datos empleados

- Se cuentan con dos bases de datos:
  - Base Icfes: incluye datos de las pruebas Saber11, familiares, socioeconómicos e información del colegio de origen.
  - Base Universidad de los Andes: incluye datos académicos (notas parciales y finales, número de créditos) y datos socioeconómicos.
- La base de datos final contiene información para el período 2015-2016/2017.
- Abarca las siguientes asignaturas:
  - Cálculo Diferencial.
  - Física I y II.
  - Química I.
  - Algebra Lineal.
- La marca de aprobación fue construida como sigue:

$$y_i = \begin{cases} Aprobado : Final \geq 3.0 & y = 1 \\ Reprobado : Final < 3.0 & y = 0 \end{cases} \quad (1)$$

Datos Icfes

Cuadro 1: Variables

Tipo de variable	Variable	Descripción
Geográficas	Departamento	
	Municipio	
Económicas	Valor pensión colegio	5 categorías
	Ingreso familiar mensual	7 categorías
Saber 11	Veces de presentación del Exámen	4 categorías
	Puntaje Saber (Todas las áreas)	Entero
Familiares	Educación madre/padre	11 categorías
	Ocupación madre/padre	12 categorías
	Número de personas en el hogar	12 categorías
	Cuartos del hogar	9 categorías
Socio-Demográficas	Estrato	6 categorías
	Nivel sisben	5 categorías
	Género	2 categorías
	Edad	Entero
Colegio	Naturaleza	2 categorías
	Carácter	4 categorías
	Área	2 categorías
Otras	Internet	Binaria
	Computador	Binaria

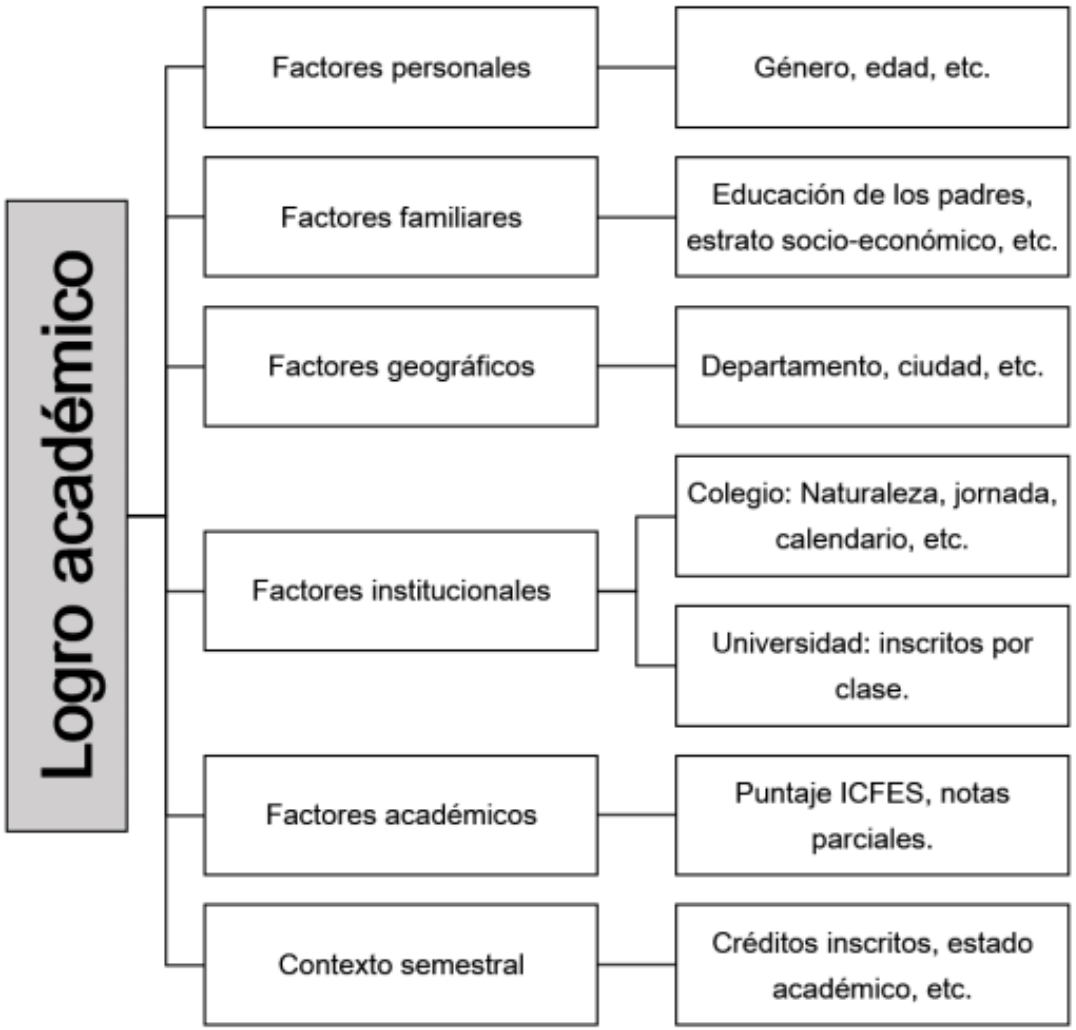
Datos Uniandes

Cuadro 2: Variables

Tipo de variable	Variable	Descripción
Socio-Demográficas	Estrato	6 categorías
	Género	2 categorías
	Edad	Entero
Colegio	Ubicación y Nombre	
Académicas	Notas parciales y finales	Real
	Promedio semestral y acumulado	Real
	Número de créditos	Entero
	Facultad, programa	Categórica
	Clasificación en Matemáticas	Entero
	Financiamiento	Categórica
	Situación académica	Categórica

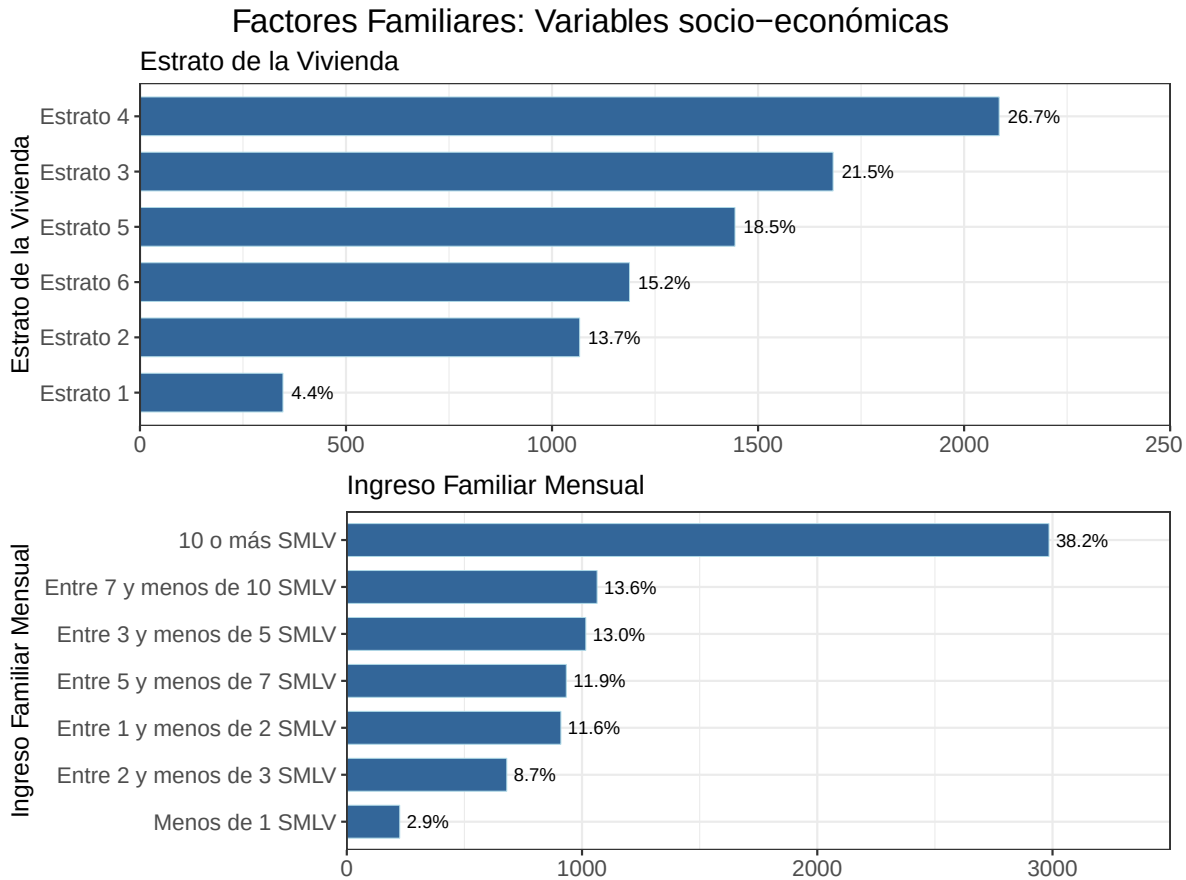
Factores Explicativos

Figura 1: Factores



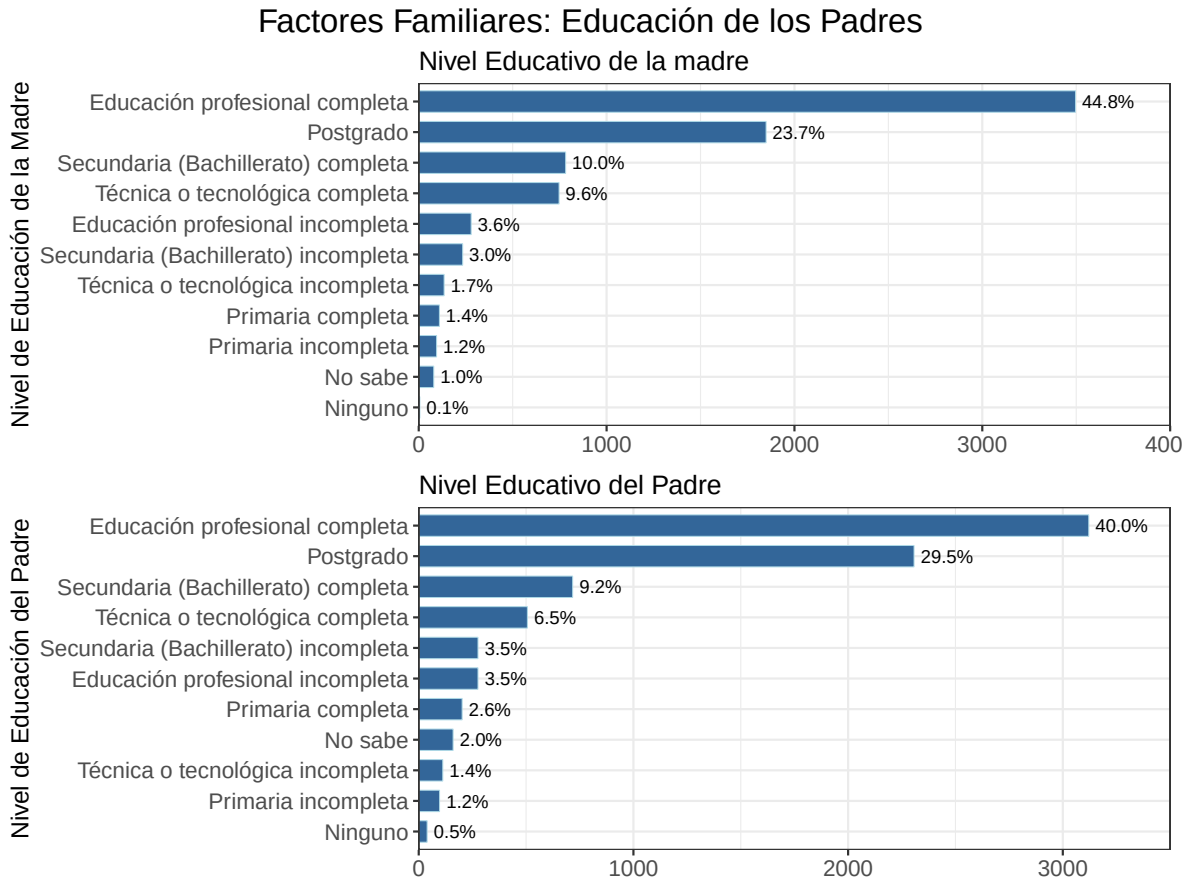
Descripción de algunas variables

Figura 2: Variables Socioeconómicas



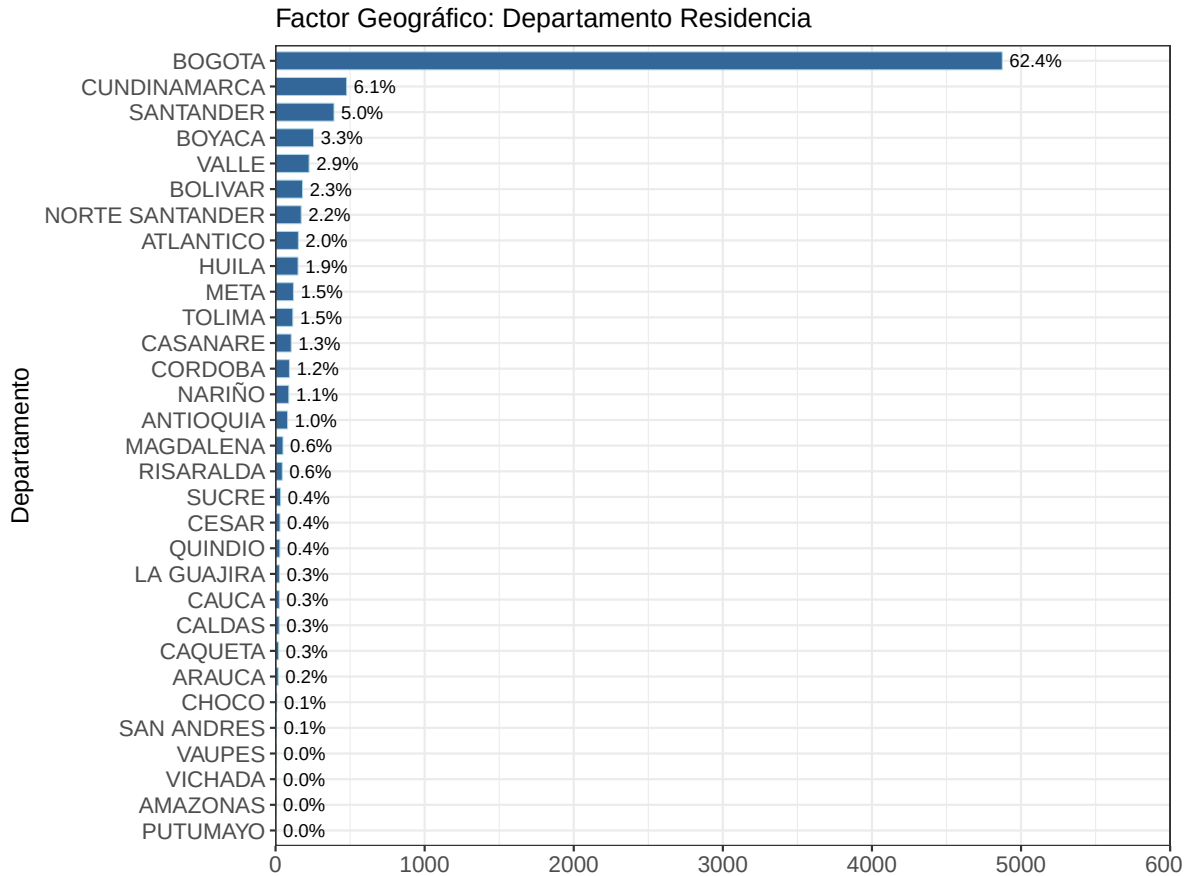
Descripción de algunas variables

Figura 3: Variables Familiares



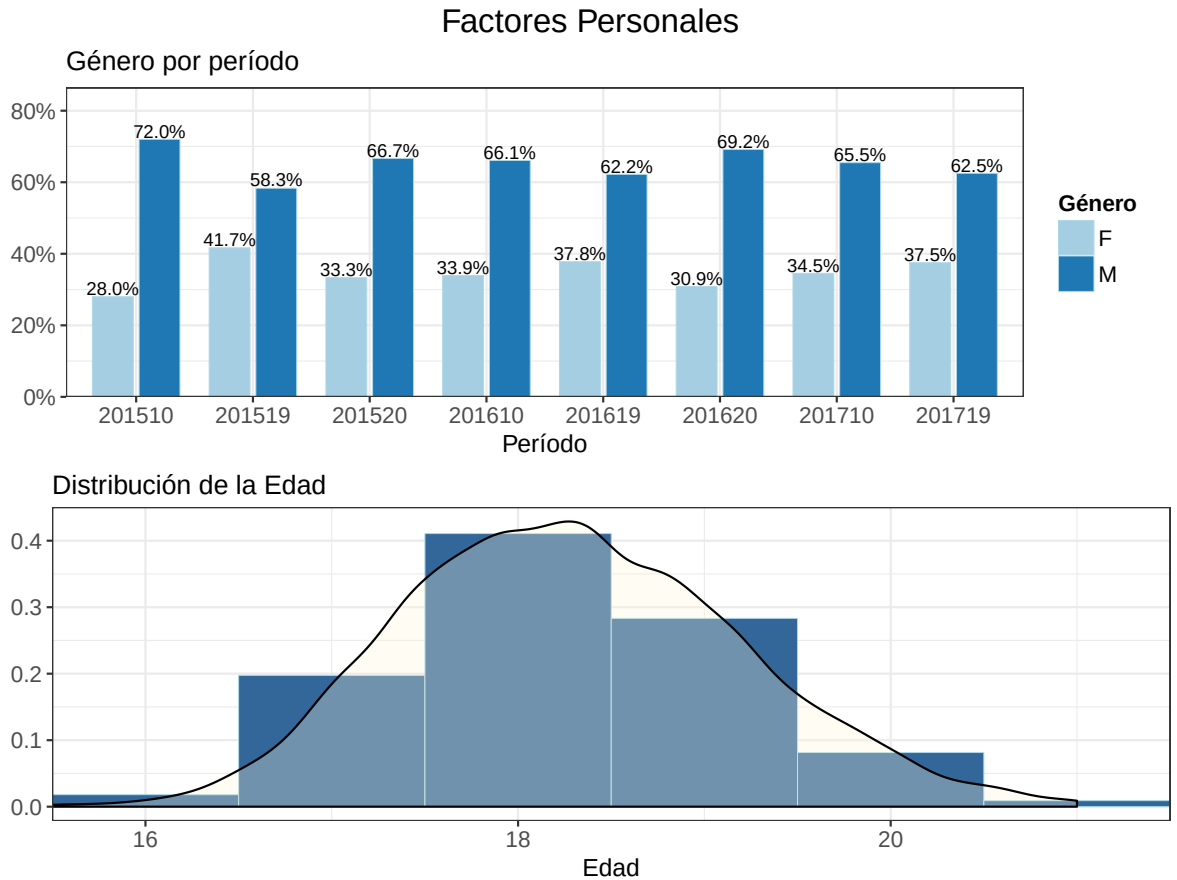
Descripción de algunas variables

Figura 4: Variable Geográfica



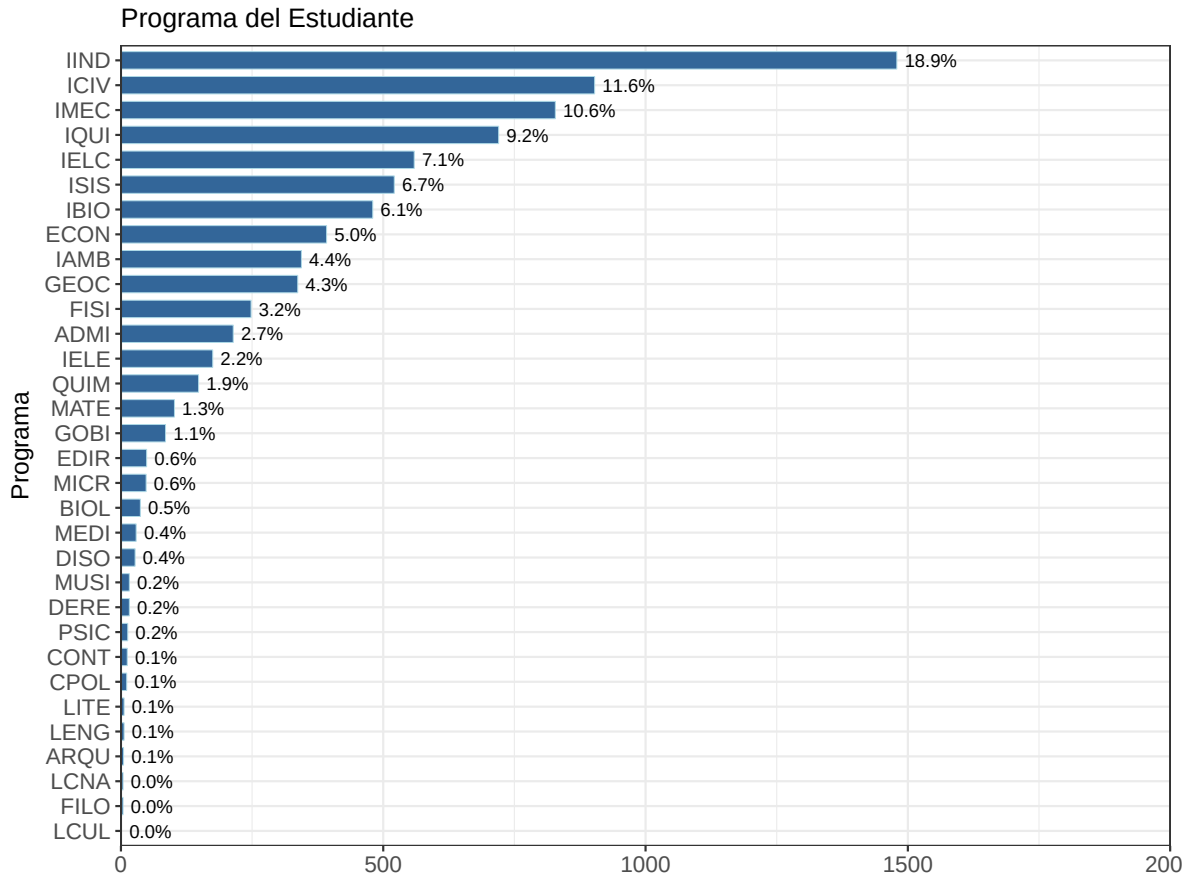
Descripción de algunas variables

Figura 5: Variable Personal



Descripción de algunas variables

Figura 6: Programa



Metodología

- Se construyen tres tipos de modelos:
  1. Ex-Ante: No incluye notas parciales de la asignatura.
  2. P1: Incluye sólo nota parcial 1.
  3. Ex-Post: Incluye notas parciales 1 y 2.
- Para el modelo Ex-Ante se cuenta con 7,823.
- Para los modelos P1 y P2 se cuentan con 3,660 observaciones[1].
- Se transforman las variables en pesos de evidencia.
- Se estiman cinco tipos de modelo:
  - Regresión Logística (*Stepwise*).
  - Regresión Lasso.
  - *Boosting* de árboles.
  - *Support Vector Machine*.
  - Redes Neuronales.

[1] El cruce de información se realiza principalmente con el documento de identificación que existe en la bases de datos de la Uniandes sólo se recupera alrededor del 33 % de los id's enviados. Adicionalmente, existen observaciones que no figuran con notas parciales, sólo con nota definitiva.

Técnicas de estimación

1. Regresión Logística:
   - Estima la probabilidad de que una característica esté presente (probabilidad de éxito) dada un conjunto de variables explicativas.

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \tag{2}$$

$$l(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] \tag{3}$$

2. Regresión Lasso:
   - La regresión Lasso es un tipo de regresión que hace uso de regularización, en particular impone una penalización L1 a los coeficientes que más contribuyen al error.

$$l_\lambda^R(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{4}$$

Transformación de los datos

- Pesos de Evidencia (WOE, por sus siglas en inglés) → Describe relación entre una variable predictora y una variable objetivo (Categorización o transformación).

Formalmente, si $B_1, \ldots, B_k$ son las categorías de X_j, el WoE de X_j para cada categoría i puede denotarse como:

$$\text{WoE}_{ij} = \log \frac{P(X_j \in B_i | Y = 1)}{P(X_j \in B_i | Y = 0)} = \log \frac{\textit{Distrib. de Incumplidos}_i}{\textit{Distrib. de Cumplidos}_i},$$

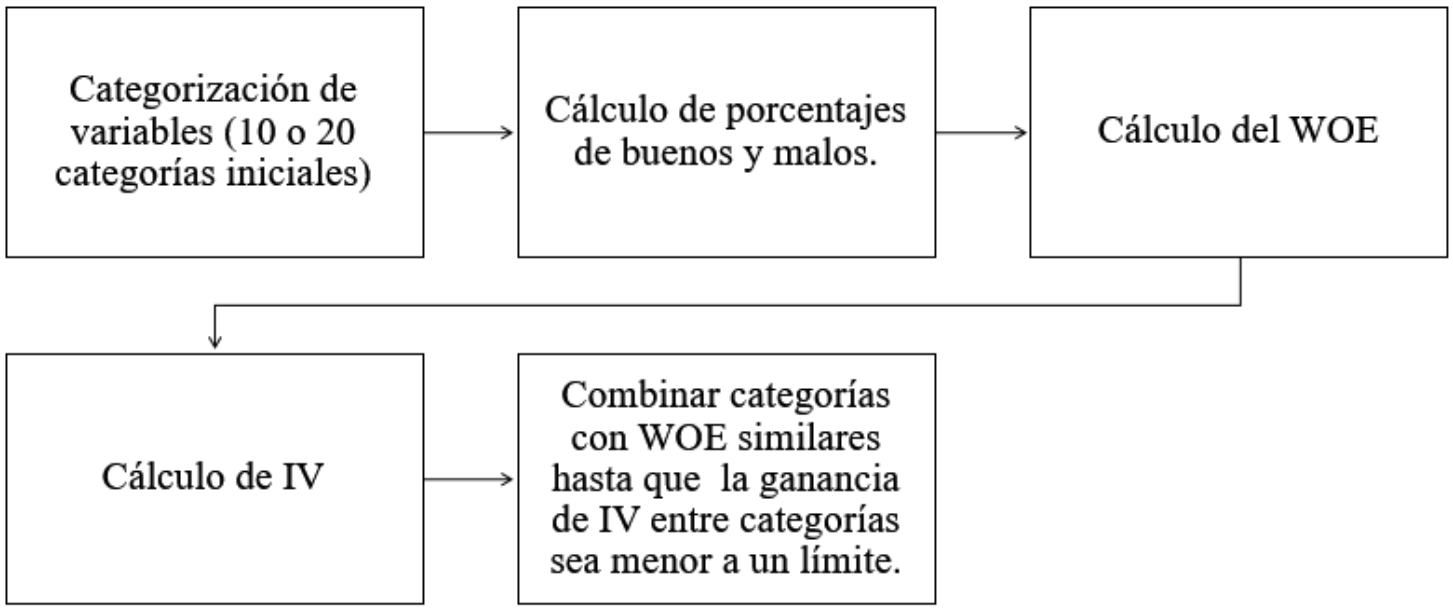
$$\text{IV}_j = \sum_{i=1}^{k} (P(X_j \in B_i | Y = 1) - P(X_j \in B_i | Y = 0)) \times \text{WOE}_{ij}$$

$$= \sum_{i=1}^{k} (\textit{Distrib. de Incumplidos}_i - \textit{Distrib. de Cumplidos}_i) \times \text{WOE}_{ij},$$

Transformación de los datos

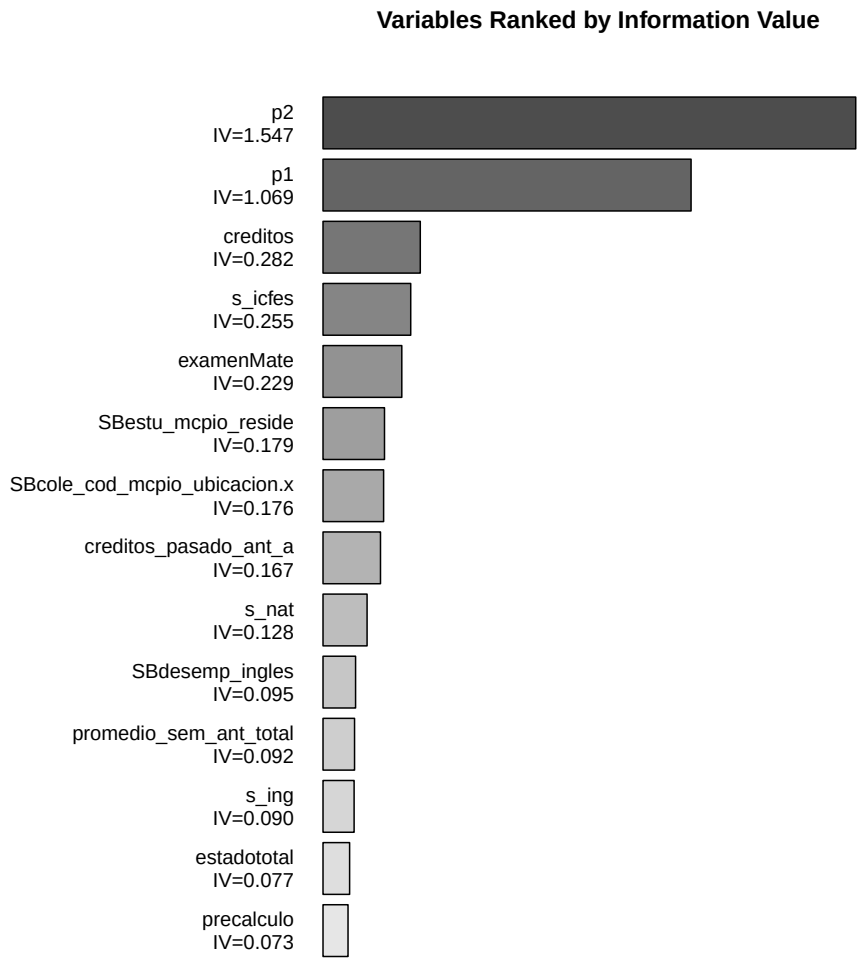
Proceso de categorización y pre-selección de variables con IV y WOE:

Figura 7: Proceso de Transformación



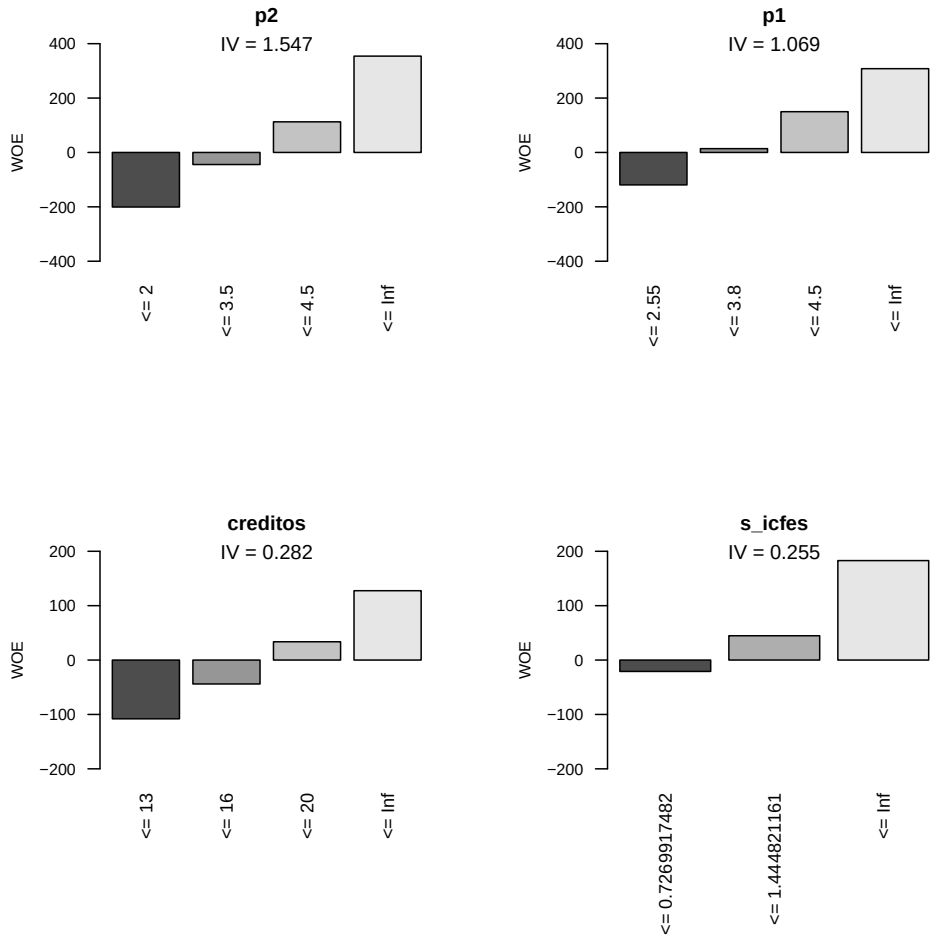
Transformación de los datos: Modelo Ex-Post

Figura 8: Ranking Variables para Modelo Ex-Post



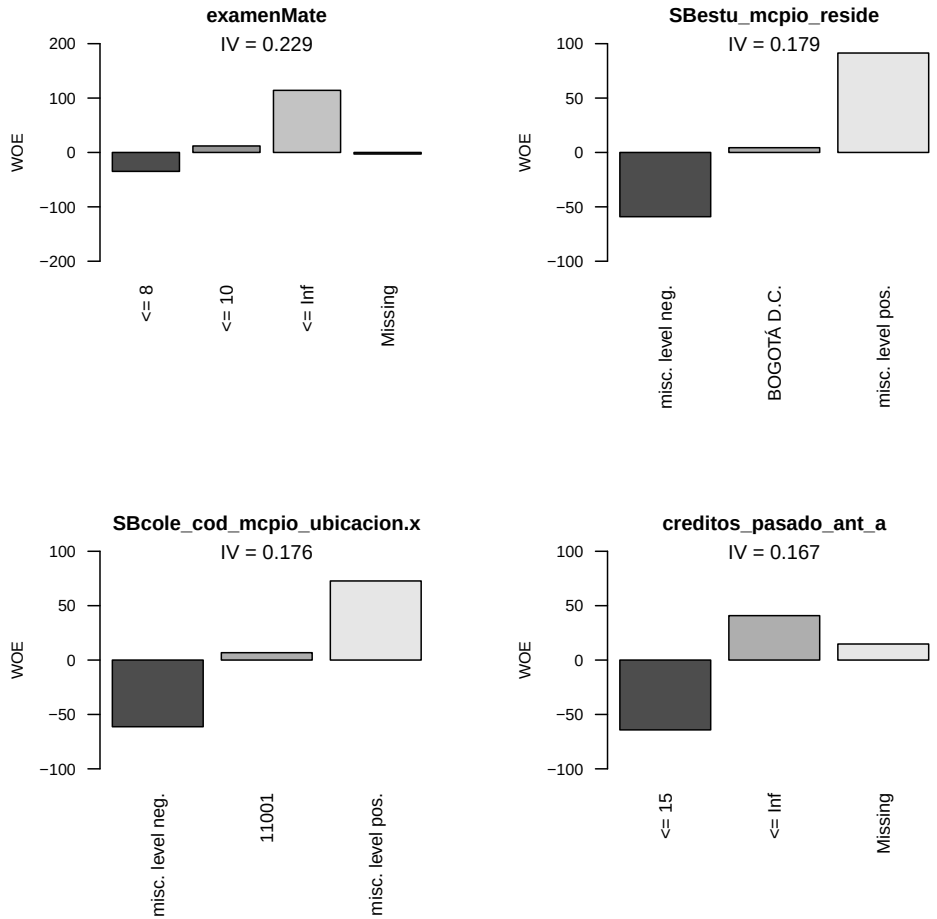
Transformación de los datos: Modelo Ex-Post

Figura 9: WOE-Modelo Ex-Post



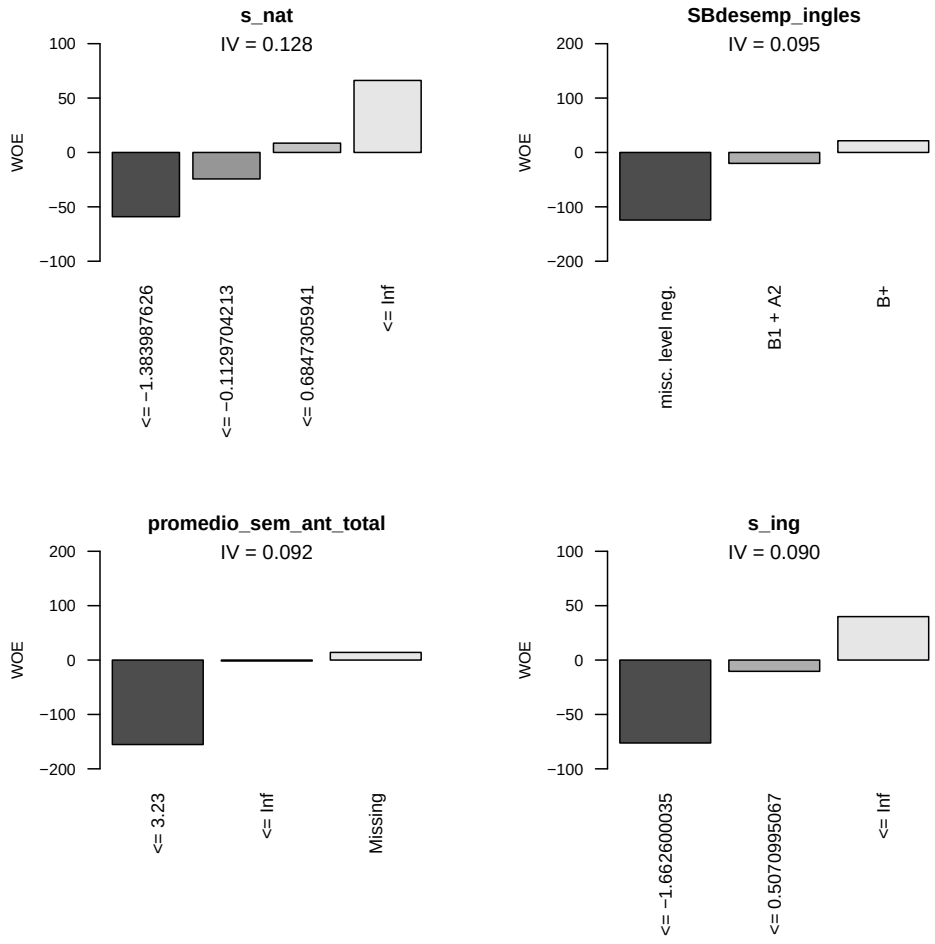
Transformación de los datos: Modelo Ex-Post

Figura 10: WOE-Modelo Ex-Post



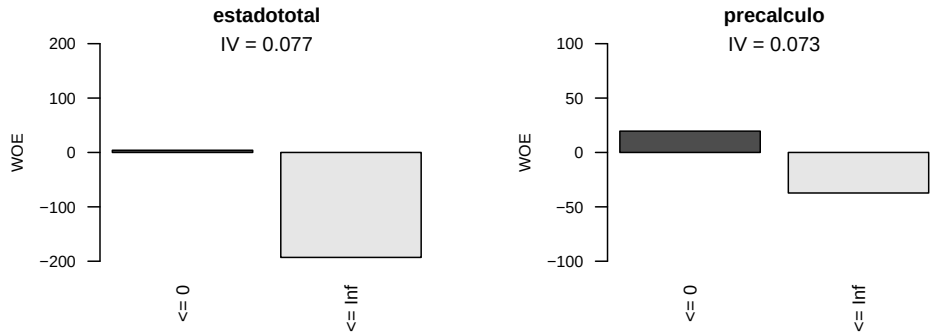
Transformación de los datos: Modelo Ex-Post

Figura 11: WOE-Modelo Ex-Post



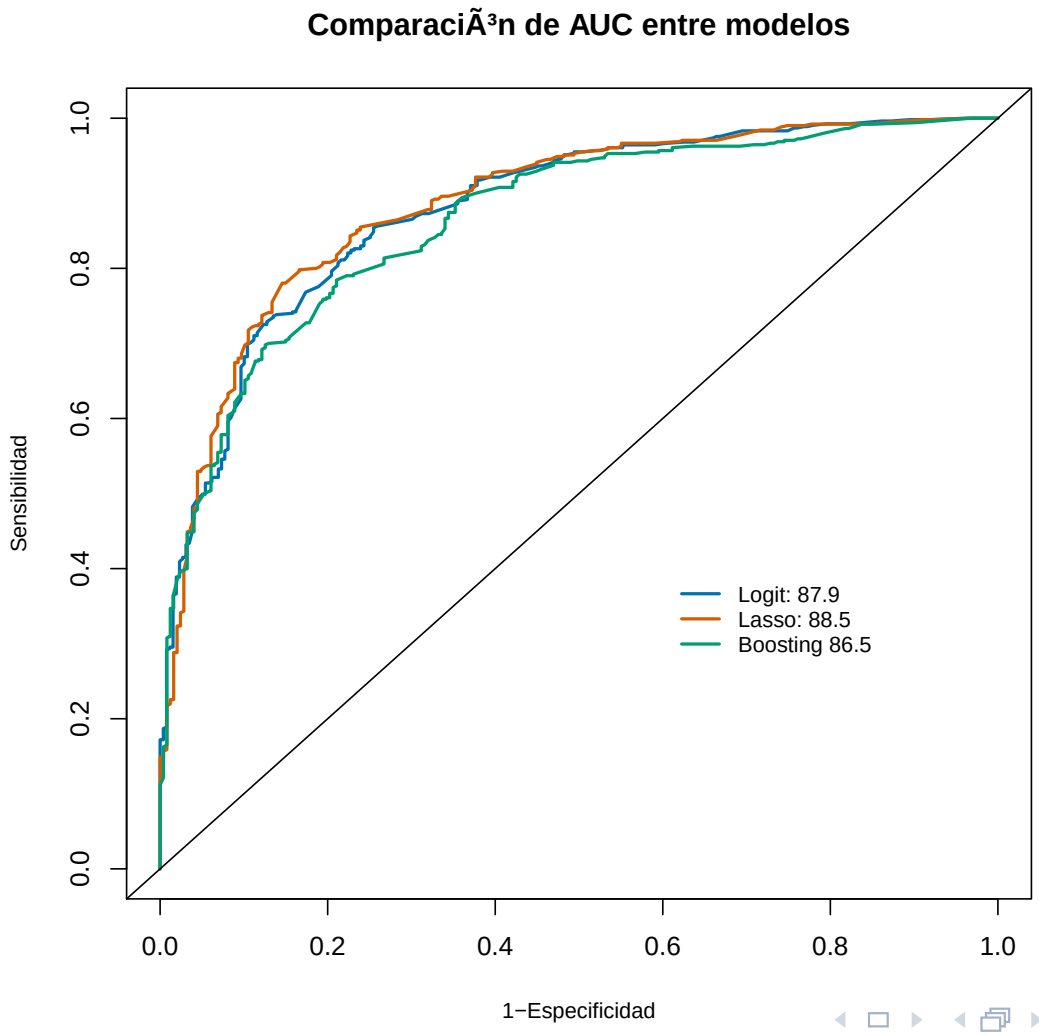
Transformación de los datos: Modelo Ex-Post

Figura 12: WOE-Modelo Ex-Post



Resultados: Modelos Ex-Post

Figura 13: Comparación Modelos Ex-Post



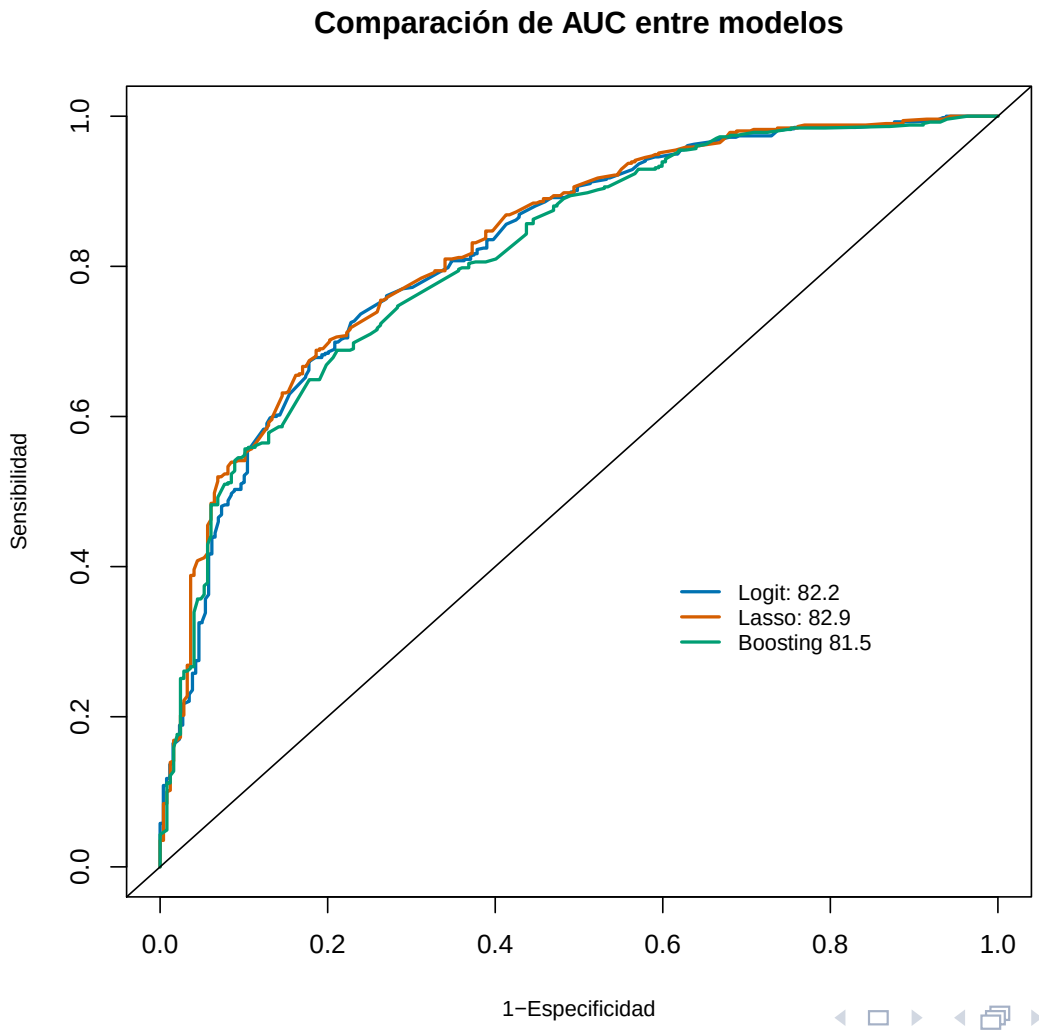
Variables por Modelo Ex-Post

Cuadro 3: Variables Por Modelo

	Logit	Lasso	Boosting
1	woe.p2.binned	woe.p2.binned	woe.p2.binned
2	woe.p1.binned	woe.p1.binned	woe.p1.binned
3	woe.estadototal.binned	woe.creditos_pasado_ant_a.binned	woe.SBestu_mcpio_reside.binned
4	woe.SBestu_mcpio_reside.binned	woe.SBestu_mcpio_reside.binned	woe.creditos_pasado_ant_a.binned
5	woe.creditos_pasado_ant_a.binned	woe.estadototal.binned	woe.examenMate.binned
6	woe.s_icfes.binned	woe.examenMate.binned	woe.estadototal.binned
7	woe.examenMate.binned	woe.s_icfes.binned	woe.s_icfes.binned

Resultados: Modelos P1

Figura 14: Comparación Modelos Ex-Post



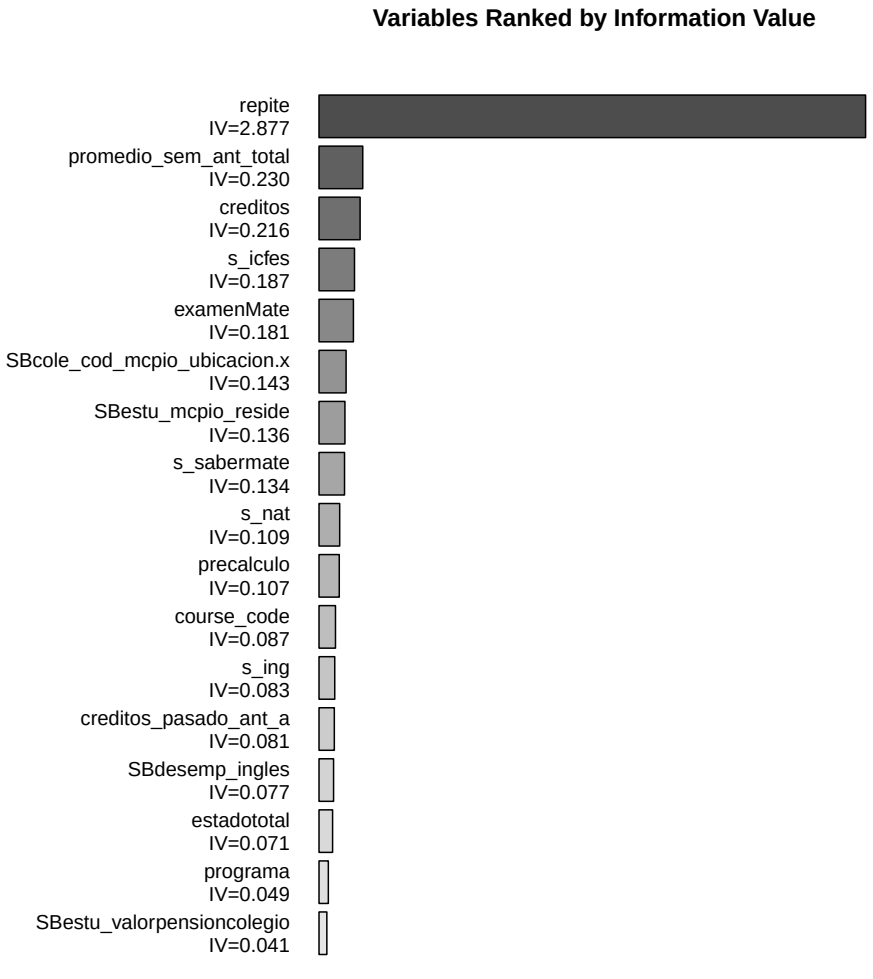
Variables por Modelo P1

Cuadro 4: Variables Por Modelo

	Logit	Lasso	Boosting
1	woe.p1.binned	woe.p1.binned	woe.p1.binned
2	woe.SBestu_mcpio_reside.binned	woe.SBestu_mcpio_reside.binned	woe.creditos.binned
3	woe.estadototal.binned	woe.creditos.binned	woe.SBestu_mcpio_reside.binned
4	woe.creditos.binned	woe.estadototal.binned	woe.creditos_pasado_ant_a.binned
5	woe.creditos_pasado_ant_a.binned	woe.creditos_pasado_ant_a.binned	woe.estadototal.binned
6	woe.s_icfes.binned	woe.s_icfes.binned	woe.s_icfes.binned
7	woe.examenMate.binned	woe.examenMate.binned	woe.examenMate.binned

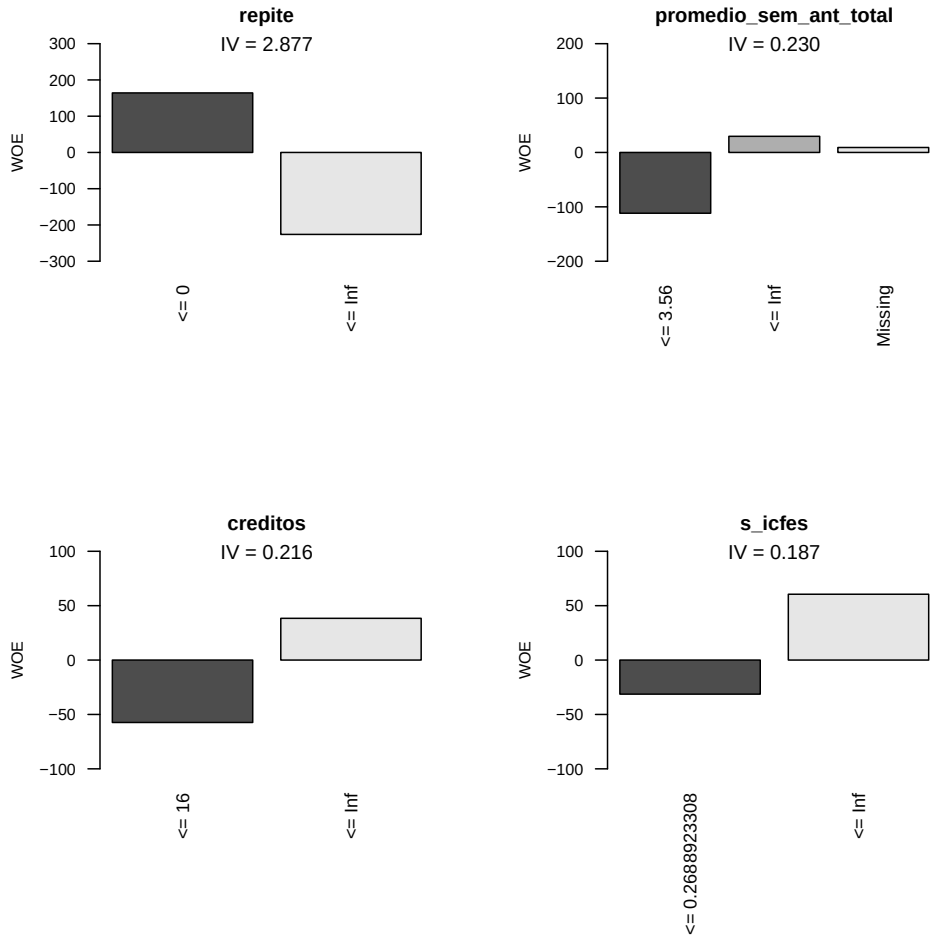
Transformación de los datos: Modelo Ex-Ante

Figura 15: Ranking Variables para Modelo Ex-Ante



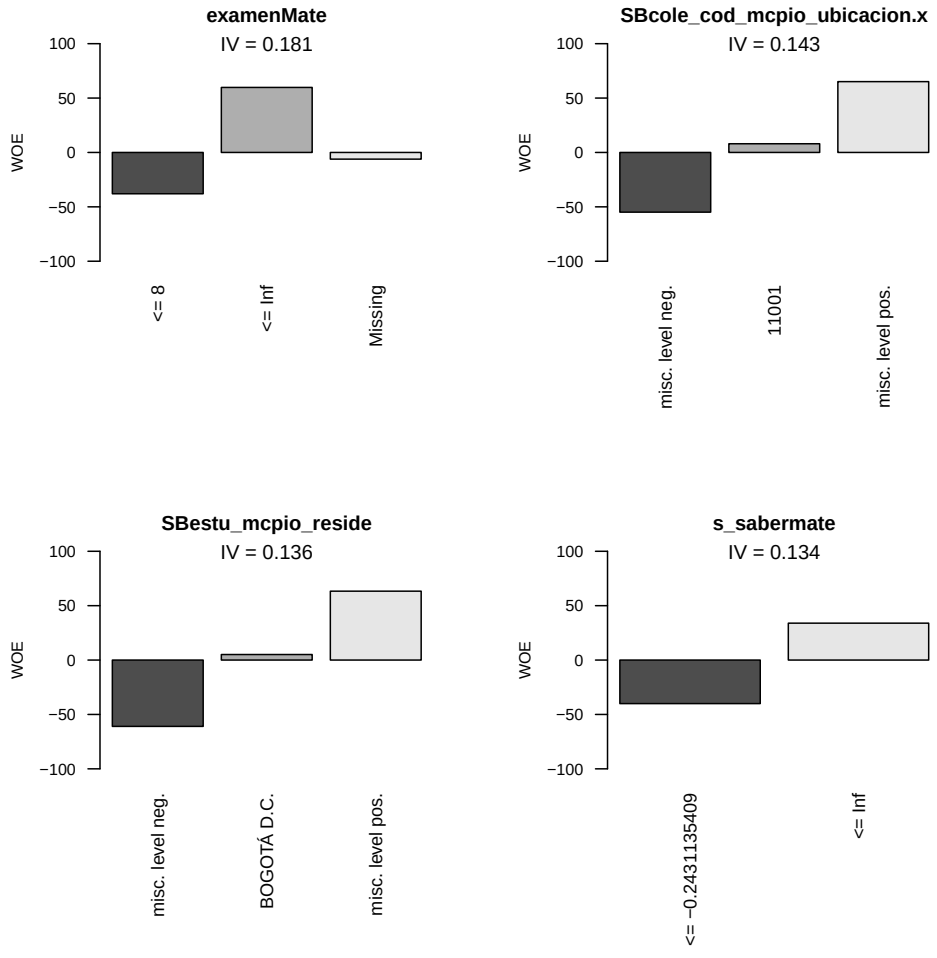
Transformación de los datos: Modelo Ex-Ante

Figura 16: WOE-Modelo Ex-Ante



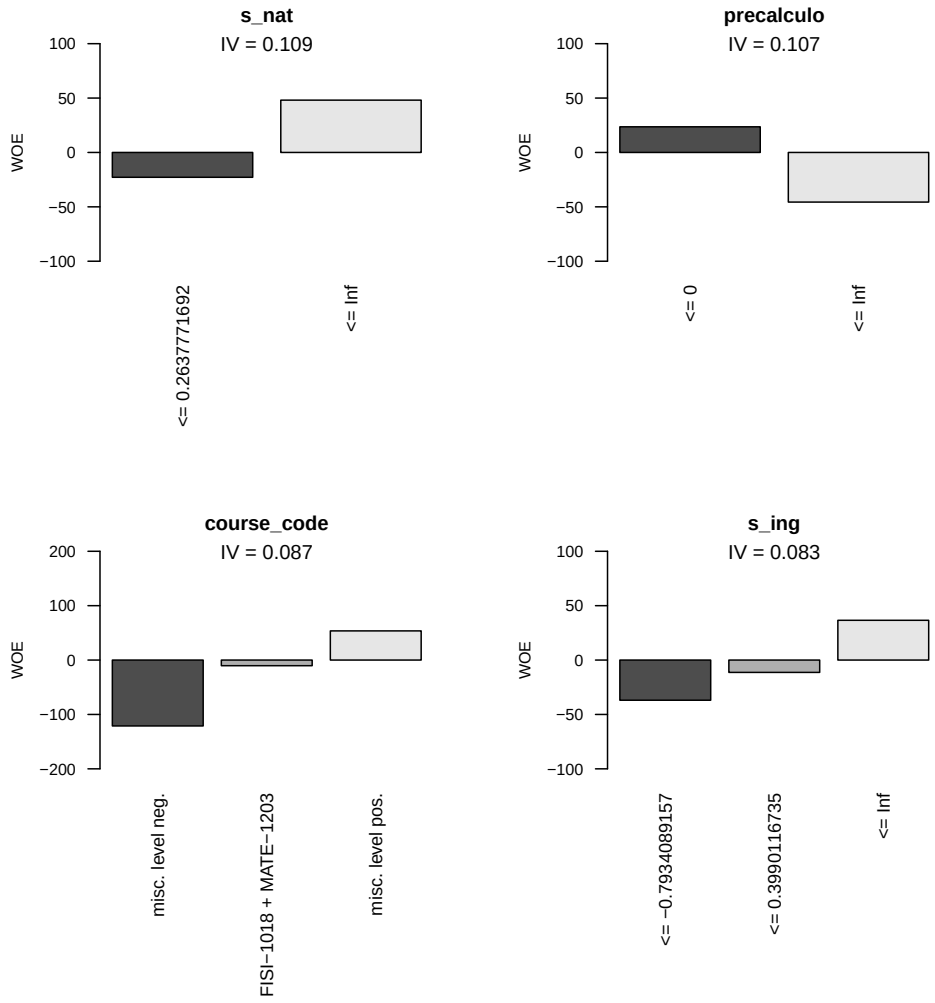
Transformación de los datos: Modelo Ex-Ante

Figura 17: WOE-Modelo Ex-Ante



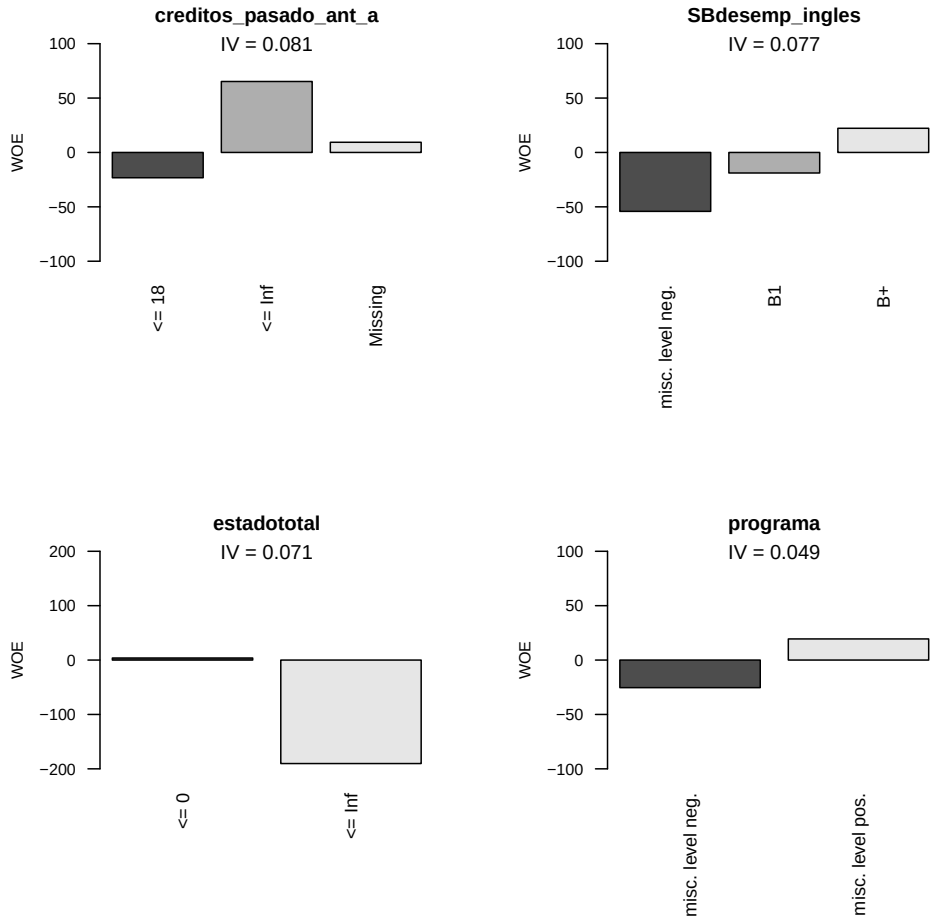
Transformación de los datos: Modelo Ex-Ante

Figura 18: WOE-Modelo Ex-Ante



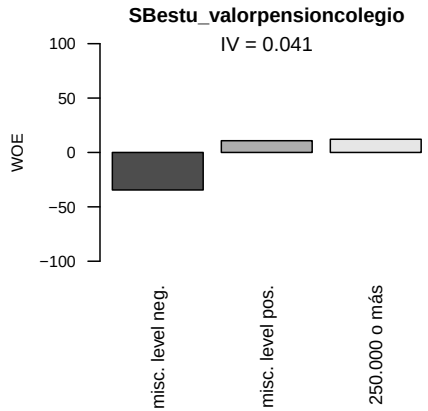
Transformación de los datos: Modelo Ex-Ante

Figura 19: WOE-Modelo Ex-Ante



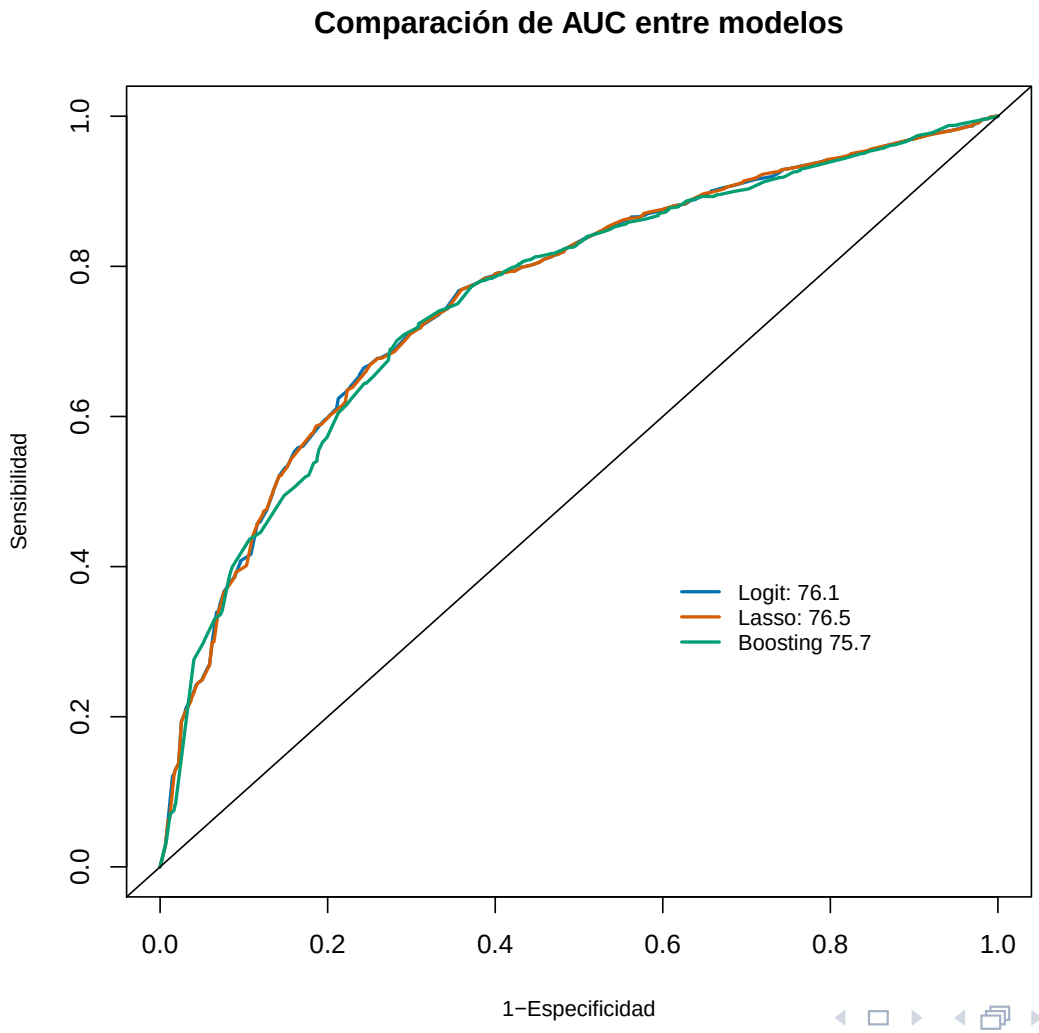
Transformación de los datos: Modelo Ex-Ante

Figura 20: WOE-Modelo Ex-Ante



Resultados: Modelos Ex-Ante

Figura 21: Comparación Modelos Ex-Ante



Variables por Modelo Ex-Ante

Cuadro 5: Variables Por Modelo

	Logit	Lasso	Boosting
1	woe.repite.binned	woe.repite.binned	woe.repite.binned
2	woe.course_code.binned	woe.course_code.binned	woe.estadototal.binned
3	woe.creditos.binned	woe.creditos.binned	woe.examenMate.binned
4	woe.estadototal.binned	woe.estadototal.binned	woe.course_code.binned
5	woe.s_sabermate.binned	woe.s_sabermate.binned	woe.creditos.binned
6	woe.examenMate.binned	woe.examenMate.binned	woe.s_sabermate.binned
7	woe.s_ing.binned	woe.s_ing.binned	woe.s_ing.binned

Próximos Pasos

- Probar nuevas formas de categorización de variables.
- Incluir variables psicométricas (DECA).
- Construcción de una base de datos más amplia.

GRACIAS