BANCO DE ESPAÑA
Eurosistema

USO INTERNO

PREDICTING THE NEED TO RECONTACT IN HOUSEHOLD SURVEY DATA: A MACHINE LEARNING APPROACH

Nicolás Forteza, Sandra García Uribe

UNIDAD DE LA ENCUESTA FINANCIERA DE LAS FAMILIAS

ÍNDICE

- **La Encuesta Financiera de las Familias (EFF) recoge información detallada sobre las rentas, los activos, las deudas y el gasto de los hogares españoles.**
- **En la primera fase de revisión de los datos recogidos de la encuesta el equipo de revisión detecta si existen casos con información incongruente, omisiones y errores no subsanables con la información recogida durante la entrevista.**
- **En una parte de estos casos se plantea *re-contactar* con los hogares para preguntar nuevamente sobre algunas partes clave de la encuesta, con el fin de evitar el descarte de cuestionarios en su totalidad, y mantener la representatividad de la muestra y calidad de los datos finales.**
- **La detección de cuestionarios a *re-contactar* es una tarea que se viene haciendo de forma manual y que requiere coordinación en los criterios de revisión dentro del equipo.**
- **Actualmente los casos se revisan a medida que se completa el trabajo de campo, y no existe ningún criterio de priorización.**

- El objetivo de este estudio es encontrar un modelo estadístico de clasificación que sea capaz de predecir si existe necesidad de re-contactar a un hogar.
- El output del modelo deberá ser un score o métrica que cuantifique la necesidad de re-contactar.
- Se aplican técnicas de *machine learning* sobre datos ya clasificados de oleadas anteriores para ajustar el modelo y encontrar el modelo óptimo, evaluando los resultados en un set de testeo.
- Organizamos datos de la encuesta de distinta naturaleza, que a veces se presentan de forma desestructurada: paradata, contadores, comentarios textuales del entrevistador y variables del cuestionario.
- Encontramos un algoritmo y un conjunto de datos que son capaces de generar un score relativamente preciso.
- Encontramos una serie de variables que ayudan a explicar en mejor medida el output del modelo.
- Encontramos una coherencia en la interpretabilidad del modelo.

- **A. Kennickell (2006) ya menciona la importancia de detectar casos de alta prioridad.**

- **Otras disciplinas y casos de uso de éxito donde se aplican técnicas similares:**
  - **Riesgo de crédito (Alonso y Carbó, 2020; Khandani et al., 2010).**
  - **Detección de fraude (Ngai et al., 2011; Phua et al., 2010).**
  - **Abandono o *churn* (Hung et al., 2006).**

- **La metodología presentada en este paper pretender complementar la literatura relacionada con *survey data editing* (Little et al., 1984; Ghosh-Dastidar et al., 2006, de Waal, 2005).**

- **Variable dependiente: "RE-CONTACTO".**
  - Según el trabajo de campo se va realizando, el equipo de revisión analiza exhaustivamente los cuestionarios, con el fin de detectar errores.

$$Y = \{y_1, y_{2,\ldots,} y_i\}; y \in \{0, 1\}$$

0: No recontacto

1: Recontacto

- **Cuestionario: información reportada por el hogar en bruto (sin editar) de la encuesta.**
  - **Información básica del hogar:**
    - Número de miembros, uso de proxy en la entrevista, si es un hogar de tipo panel, contadores (número de preguntas contestadas según su naturaleza) y ratios de no respuesta.
  - **Algunas variables reportadas por hogares:**
    - Género de persona responsable (PR), edad PR, nivel de satisfacción con su vida, nivel de educación de PR.

- Activos reales: régimen de tenencia de vivienda principal, número de propiedades, número de negocios y sus características (cómo lo obtuvieron, uso de avales).
- Activos financieros: si poseen acciones, valores de renta fija, fondos de inversión, etc.
- Planes de pensiones.

- **En función de las variables anteriormente mencionadas,** se computan diversos indicadores que pueden indicar sospecha de cara a revisar un cuestionario:
  - Duplicidad en el reporte de los beneficios de un negocio.
  - Discrepancia en la fecha de percepción de las rentas de pensiones.
  - Confusión en situación laboral (negocio vs. autónomo).
  - Omisión de situación laboral.
  - Infraestimación de ingresos por renta por cuenta propia.

- **Variables de entrevistador:**
  - Son preguntas que rellena el entrevistador al finalizar la entrevistar.
  - Se dan variables también de comentarios.
- **Comentarios a lo largo de la entrevista.** Los entrevistadores pueden realizar anotaciones mientras realizan la entrevista con el fin de aclarar la información reportada por un hogar.

- **Dos fuentes de información diferentes de comentarios de entrevistador:**
  - Sección final de entrevista.
  - Los realizados a lo largo de la entrevista.
- **Después de aplicar un pipeline de limpieza de los comentarios, para cada conjunto de datos, se extraen diferentes features:**

1. Longitud de comentarios

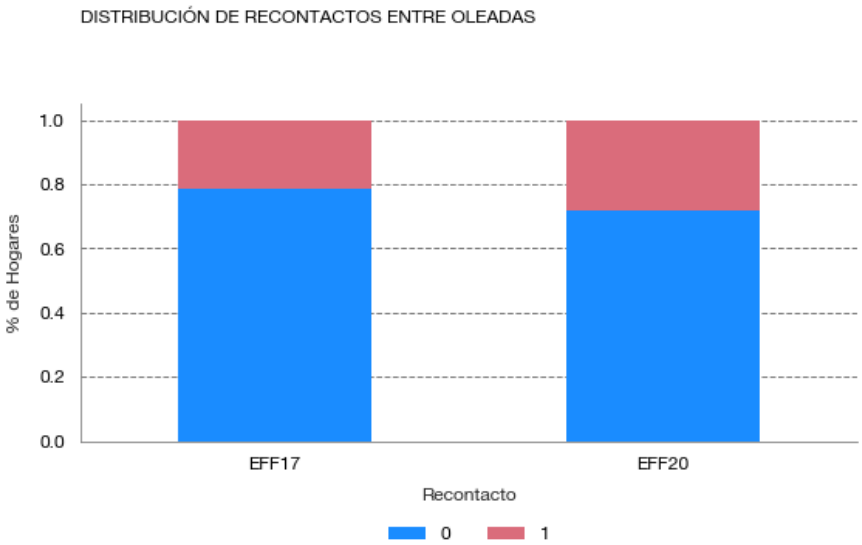
1. Métrica TF-IDF (Term frequency – Inverse Document Frequency), ignorando términos que aparecen en menos del 5% de cuestionarios:

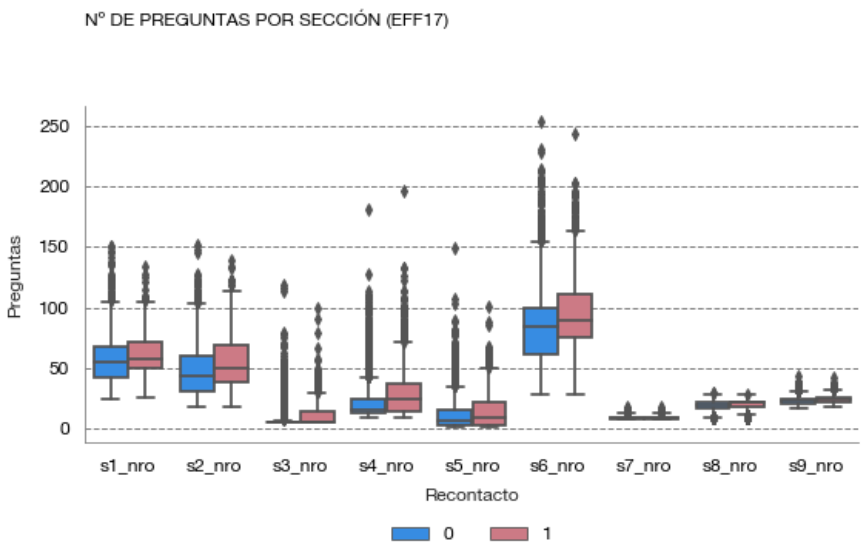
$$TF - IDF = tf(t,d) * \log\left(\frac{n}{1 + df(t)}\right)$$

2. SVD (*singular value decomposition*) de los primeros 50 componentes sobre una matriz TF-IDF sin mínimo de frecuencia.

- **Timing: variables relativas a la hora de realización de la entrevista.**
  - Día laboral vs fin de semana, franja horaria del día, días desde comienzo del trabajo de campo.
- **Paradata: información relativa a las duraciones de la entrevista.**
  - Tiempo total (en segundos) de duración de la sección i, tiempo total cuando se preguntan n categorías, total de repeticiones de alguna pregunta en sección i, total de veces que el entrevistador da a "volver" en CAPI en sección i, número de preguntas únicas formuladas (no aplican repetidas).
- **Información del entrevistador: nivel de educación, edad, experiencia y scoring en formación interna.**

- **Tamaño de la muestra: ~7900 hogares (6400 casos de EFF17 y 1500 de EFF20)**



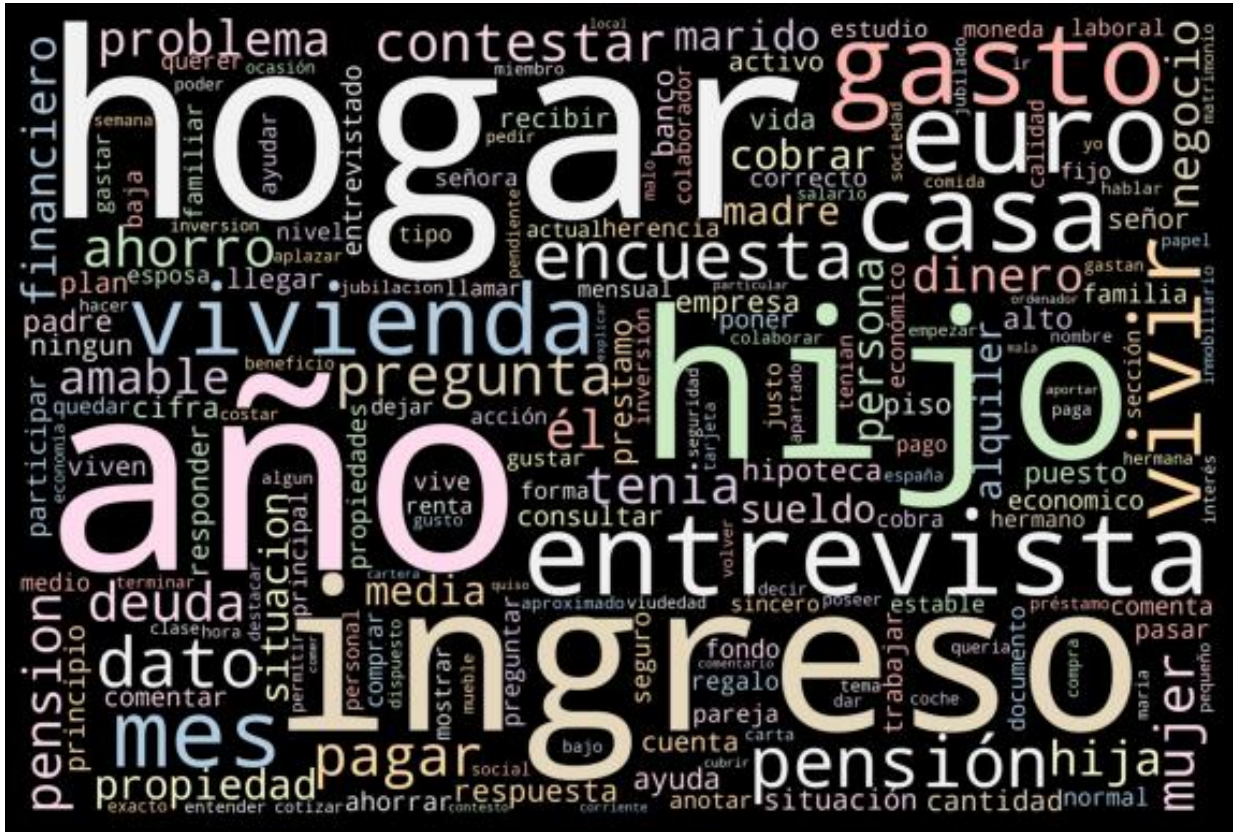


Dada la cantidad de recursos en la actual oleada, se han identificado ligeramente más recontactos que en la ola anterior. Esta misma casuística nos hizo descartar oleadas anteriores a 2017.

El promedio y la mediana del número de preguntas formuladas en los casos que fueron clasificados como re-contacto en 2017 es superior a los no clasificados como re-contacto.

- ***Wordcloud*** **de comentarios a lo largo de la entrevista.**

- ***Wordcloud*** **de comentarios de sección de entrevistador.**



- **Dados los datos de los que disponemos, encontrar una función o modelo que sea capaz de predecir Y; i.e. encontrar una función que minimice el error ε y que sea capaz de generalizar para las oleadas posteriores de la encuesta.**

$$Y = f(X, \varepsilon)$$

- **Experimentación con 5 algoritmos clasificadores:**
  - **K-vecinos.**
  - **Naive Bayes Classifier.**
  - **Regresión Logística con penalización L1.**
  - **Random Forest.**
  - **Extreme Gradient Boosting.**

- **En un primer momento, optamos por utilizar técnicas de resampling, pero con el descarte de la oleada de 2014, el resampleo no introducía efectos positivos.**

- **Para estudiar la sensibilidad al uso de diferentes subconjuntos de datos, se ajusta cada clasificador con diferentes sets de datos:**
  - **Datos de solamente la encuesta.**
  - **Datos de la encuesta + Comentarios.**
  - **Comentarios + Paradata.**
  - **Encuesta + Paradata.**
  - **Todos en conjunto.**

- **Dado el limitado tamaño de la muestra total y la no convergencia del error, se decide usar un 90% de la muestra para entrenar el modelo.**

- **Para cada algoritmo clasificador, se realiza una búsqueda de hiperparámetros con validación cruzada (3-fold stratified sampling).**

- **A la hora de elegir los hiperparámetros (en la evaluación cruzada), se maximizan varias métricas:**
  - **Área debajo de la curva ROC.**
  - **Área debajo de la curva Precision-Recall.**

- **Con los hiperparámetros óptimos, se entrenan los modelos con el 90% de la muestra de Train, y se computan las métricas en la muestra de Test con boostrap para emular el trabajo de campo (tandas de entre 100 y 300 cuestionarios).**

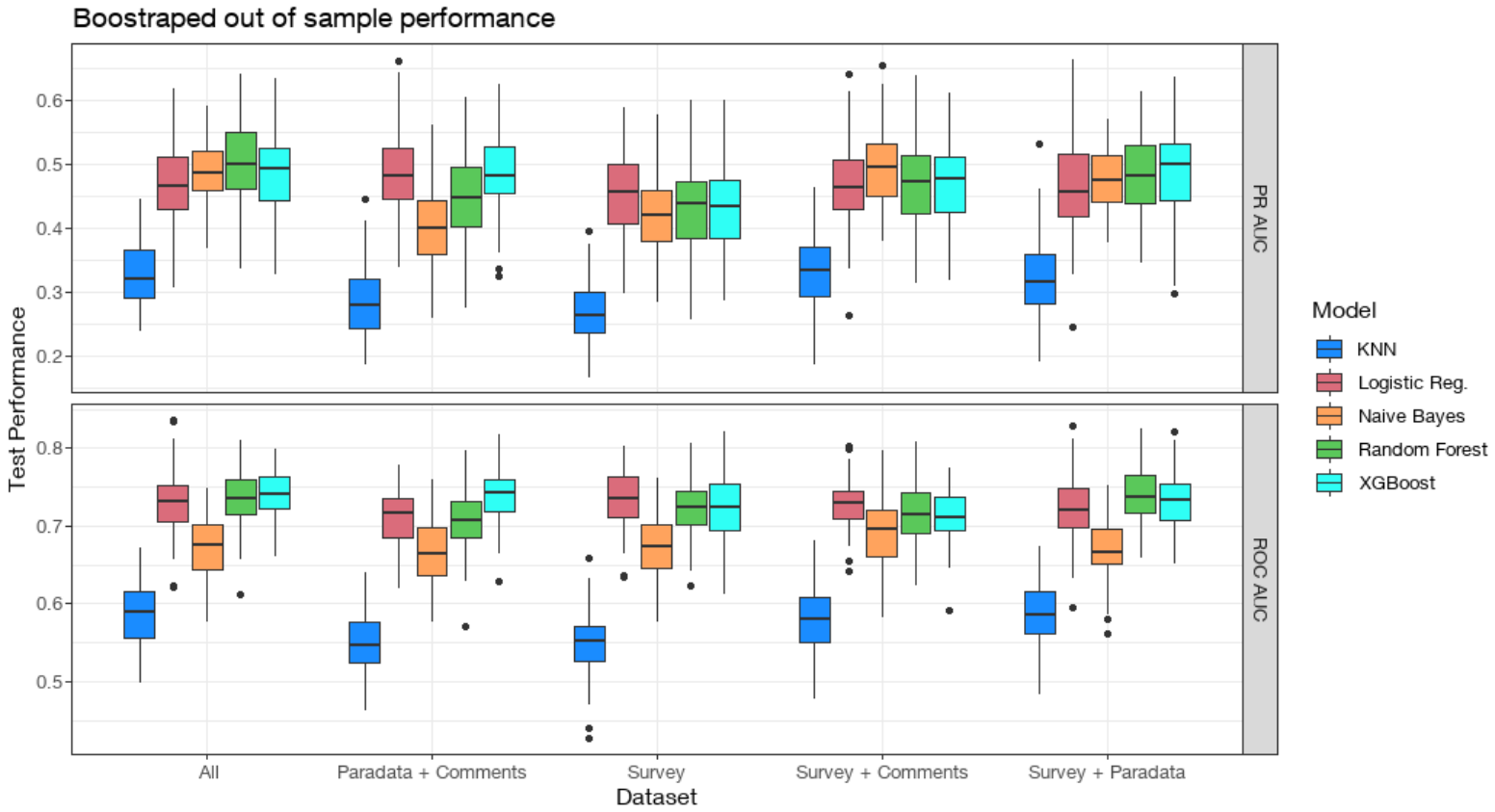
- **De cara a operativizar el caso de uso, el output de un modelo es un vector de probabilidades asociado a la muestra:**
  - **Si P >= T (umbral) → recontacto (1)**
  - **Si P < T → no recontacto (0)**

- **Dicho vector de probabilidades es el score de recontacto.**

- **Usamos una serie de métricas para evaluar el desempeño del modelo:**
  - **"Precision": de los que el modelo predice que son recontactos, cuántos en realidad lo son. (i.e.: VP / VP + FP) (i.e.: 1 – type I error)**
  - **"Recall", de los que en realidad son recontactos, cuántos he predicho correctamente. (i.e.: VP / VP + FN) (i.e.: 1 – type II error)**

- **Nuestro objetivo es elegir aquel modelo y umbral (T) en el que obtengamos un trade-off óptimo entre ambas métricas.**

- **A mayor complejidad del modelo, se dan unas mejores métricas.**
- **No existe una clara diferenciación entre los subconjuntos de datos. Usar varios subconjuntos incrementa la performance en general.**



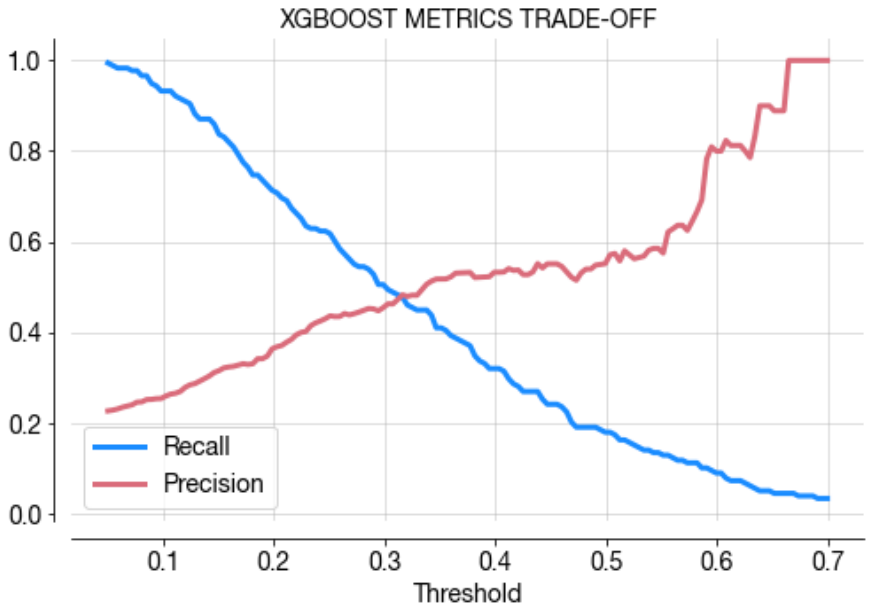
- **Existe una leve mejoría cuando se añaden las observaciones pertenecientes a esta oleada. Se espera que al final de esta oleada la curva de aprendizaje de train-test pueda converger.**

	EFF17				
	Test AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.581	0.655	0.579	0.660	0.658
Naive Bayes	0.592	0.606	0.593	0.625	0.607
Logistic Regresion (L1)	0.675	0.697	0.675	0.673	0.694
Random Forest	0.676	0.688	0.696	0.700	0.715
XGBoost	0.663	0.698	0.685	0.699	0.714

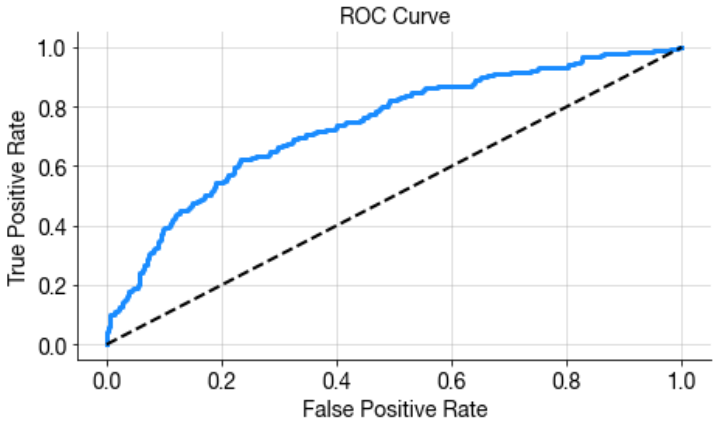
	EFF17 + EFF20				
	Test AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.630	0.660	0.643	0.632	0.658
Naive Bayes	0.673	0.662	0.687	0.669	0.674
Logistic Regresion (L1)	0.732	0.724	0.730	0.707	0.728
Random Forest	0.722	0.736	0.717	0.712	0.735
XGBoost	0.726	0.739	0.712	0.738	0.739

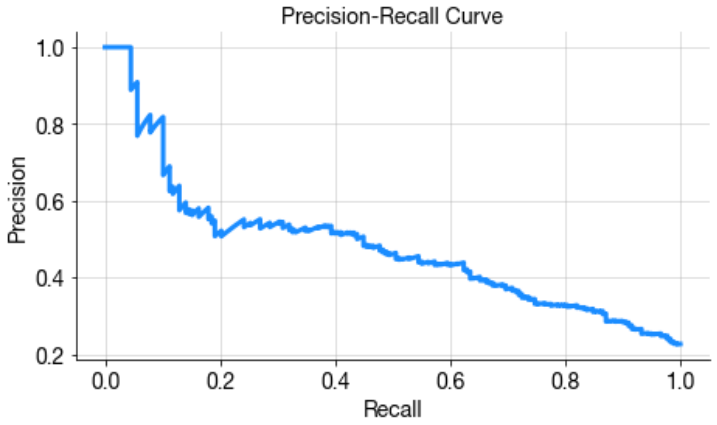
*Media de boostrapped out of simple (test) metrics

- **En general, el AUC de la curva ROC se encuentra entre 0.7 y 0.75.**
- **Destaca la buena performance de los clasificadores lineales (regresión logística) frente a los clasificadores más complejos de tipo *bagging y boosting.***
- **El trade-off entre precisión y recall es claro:**



- **Siguiendo el ejemplo anterior (XGBoost, entrenado con todos los subconjuntos de datos):**

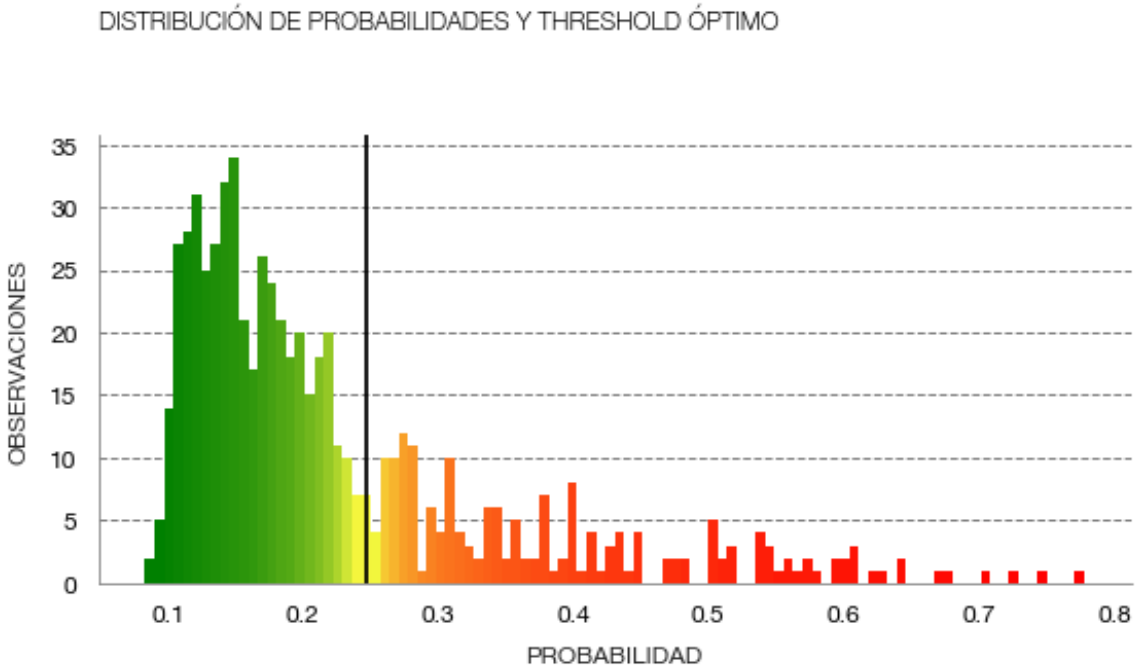




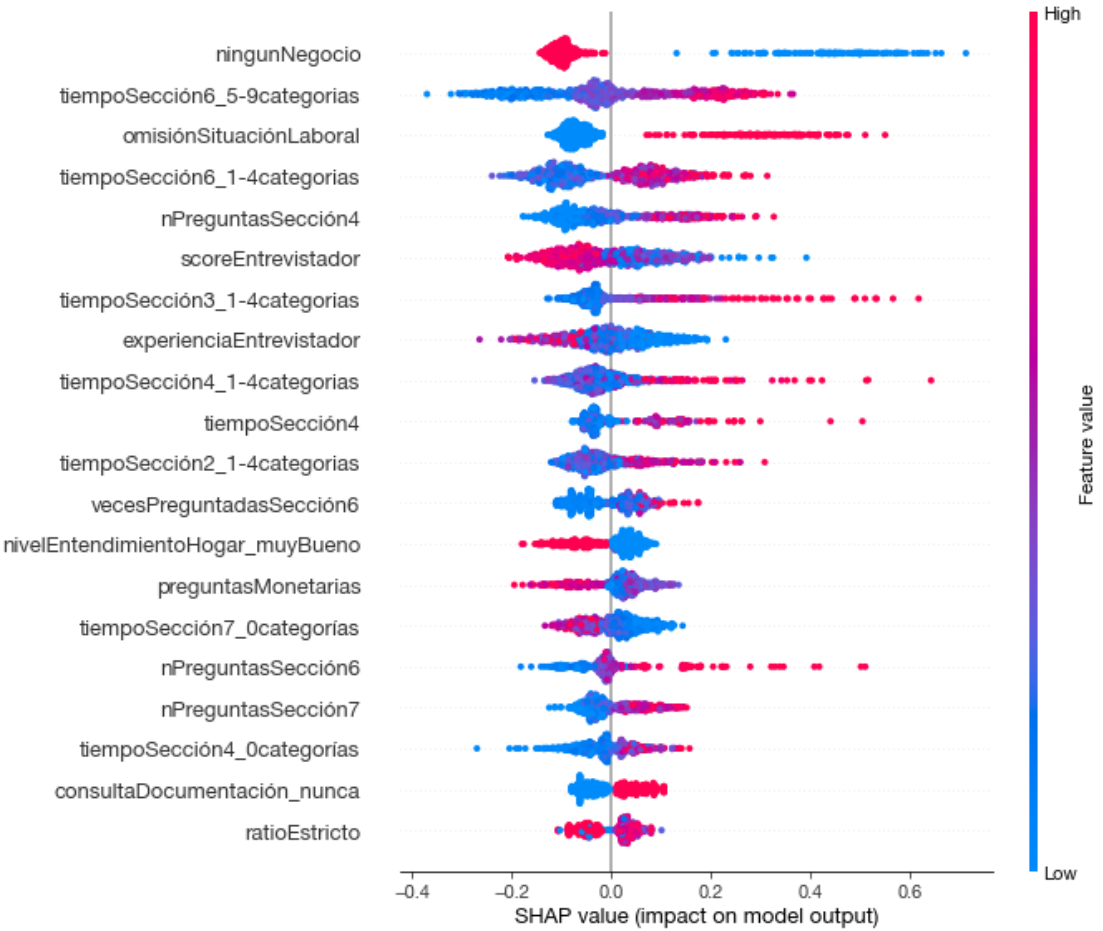
- **Podemos obtener diferentes resultados según nos vamos moviendo por las curvas:**

Threshold	0.15		0.2		0.25		0.3	
	0	1	0	1	0	1	0	1
True	288	324	393	219	470	142	506	106
False	29	149	52	126	67	111	88	90

- **El score de re-contacto (threshold = 0.25):**



- **Interpretabilidad: ¿Cuáles son los determinantes del score?**



- **Los resultados preliminares apuntan a que podemos encontrar una metodología robusta que sea capaz de generar un score de recontacto.**
- **A pesar de que existe bastante error, dicho error es asumible. La implementación de un modelo que predice el re-contacto es un gran avance respecto de las oleadas pasadas.**
- **En definitiva, el equipo de revisión puede focalizarse en un subconjunto de hogares dado el score (más o menos preciso) que indica su necesidad de ser re-contactado.**
- **Sin embargo:**
  - La naturaleza de los datos hace que sea un problema de difícil modelización.
  - Existe aún espacio de mejora en la performance de los algoritmos si:
    - *Incorporamos nuevas features que sean más discriminantes que las actuales.*
    - *Incorporamos más datos (en concreto los de la oleada actual).*

ALGUNAS REFERENCIAS

- **Kubat, M.. (2000). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. Fourteenth International Conference on Machine Learning.**
- **Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.**
- **Zheng, Zhaohui & Wu, Xiaoyun & Srihari, Rohini. (2004). Term selection for text categorization on imbalanced data. SIGKDD Explorations. 6. 80-89. 10.1145/1007730.1007741.**
- **Karakoulas et al, Optimizing Classifiers for Imbalanced Training Sets**
- **Ahmad, Hussain & Anwar, Zahid & Shah, Munam. (2017). Data mining techniques and applications — A decade review. 1-7. 10.23919/IConAC.2017.8082090.**
- **Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.**
- **A. Kennickell (2006). Who's asking? Interviewers, Their Incentives, and Data Quality in Field Surveys**

BANCO DE ESPAÑA
Eurosistema

GRACIAS POR SU ATENCIÓN

UNIDAD DE LA ENCUESTA FINANCIERA DE LAS FAMILIAS

	EFF17				
	Test AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.581	0.655	0.579	0.660	0.658
Naive Bayes	0.592	0.606	0.593	0.625	0.607
Logistic Regresion (L1)	0.675	0.697	0.675	0.673	0.694
Random Forest	0.676	0.688	0.696	0.700	0.715
XGBoost	0.663	0.698	0.685	0.699	0.714

	EFF17 + EFF20				
	Test AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.630	0.660	0.643	0.632	0.658
Naive Bayes	0.673	0.662	0.687	0.669	0.674
Logistic Regresion (L1)	0.732	0.724	0.730	0.707	0.728
Random Forest	0.722	0.736	0.717	0.712	0.735
XGBoost	0.726	0.739	0.712	0.738	0.739

	EFF17				
	Train AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.677	0.701	0.677	0.693	0.701
Naive Bayes	0.654	0.669	0.663	0.665	0.679
Logistic Regresion (L1)	0.701	0.728	0.703	0.704	0.717
Random Forest	0.941	0.994	0.914	0.980	0.961
XGBoost	0.689	0.939	1.000	0.963	0.999

	EFF17 + EFF20				
	Train AUC ROC				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.689	0.699	0.694	0.677	0.706
Naive Bayes	0.663	0.667	0.688	0.673	0.686
Logistic Regresion (L1)	0.710	0.713	0.716	0.704	0.735
Random Forest	0.881	0.954	0.852	0.850	0.853
XGBoost	0.706	0.725	0.787	0.779	0.940

	EFF17				
	Test AUC PR				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.268	0.355	0.268	0.377	0.356
Naive Bayes	0.354	0.387	0.402	0.375	0.410
Logistic Regresion (L1)	0.338	0.409	0.336	0.406	0.412
Random Forest	0.358	0.422	0.392	0.425	0.438
XGBoost	0.351	0.409	0.385	0.373	0.419

	EFF17 + EFF20				
	Test AUC PR				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.348	0.404	0.388	0.329	0.412
Naive Bayes	0.427	0.466	0.484	0.403	0.483
Logistic Regresion (L1)	0.444	0.455	0.461	0.479	0.464
Random Forest	0.435	0.479	0.462	0.454	0.494
XGBoost	0.433	0.495	0.467	0.469	0.480

	EFF17				
	Train AUC PR				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.356	0.401	0.355	0.392	0.400
Naive Bayes	0.386	0.398	0.439	0.370	0.418
Logistic Regresion (L1)	0.388	0.429	0.389	0.402	0.417
Random Forest	0.826	0.980	0.791	0.938	0.881
XGBoost	0.388	0.848	1.000	0.908	0.998

	EFF17 + EFF20				
	Train AUC PR				
	Survey	Survey + Paradata	Survey + Comments	Paradata + Comments	Survey + Paradata + Comments
KNN	0.385	0.396	0.393	0.378	0.415
Naive Bayes	0.410	0.426	0.439	0.383	0.448
Logistic Regresion (L1)	0.413	0.416	0.418	0.418	0.446
Random Forest	0.435	0.479	0.462	0.454	0.494
XGBoost	0.416	0.448	0.558	0.534	0.860