Para la elección del mejor modelo base para cada producto se realizó el siguiente procedimiento: preparación de datos, competición de modelos, análisis de variables importantes, calibración de parámetros y estimación del modelo final.

Para comenzar la preparación de datos se encontraron los registros inconsistentes o vacíos en cada variable. Para las variables financieras como ingresos mensuales, egresos mensuales, activos y pasivos, se imputaron para aquellos registros la mediana de la variable resultante del cálculo de la combinación entre departamento y estrato. También aquellos registros por encima del percentil 99 fueron imputados por el valor de percentil 99. Para las variables categóricas, aquellos registros donde no se encontraban con alguna categoría, se les imputo la más recurrente en las variables donde los registros desconocidos eran escasos. En el caso contrario, estos registros fueron etiquetados como una categoría adicional asignándoles el valor de \testit{desconocidos}. Después las variables categóricas fueron codificadas por valores binarios eliminando en el proceso una de las categorías. Por último, se realizó un análisis gráfico de la relación entre el evento de compra del producto $j$ $Y\_t$ y las variables predictivas que muestran las figuras ….. En el análisis podemos observar como …

Para la competición de modelos se tomaron las métricas de \textit{AUC}, \textit{Precisión} y \textit{exhaustivdad}\footnote{Las métricas son definidas en la sección \ref{metricas} en el apéndice} y tres modelos de clasificación \footnote{La explicación teórica de estos modelos se encuentra en la sección \ref{explicacion\_arboles} del apéndice.} tomando los parámetros por defecto de cada paquete estadístico \cite{h2o} y \cite{RXgboost} en el software estadístico R \cite{team2013r}. Los tres primeros modelos son \textit{logit}, \textit{Árboles de decisión} y \textit{ Árboles de clasificación con potenciación de gradiente} del paquete \cite{h2o}. .Para el modelo de árboles de decisión y el primero de potenciación de gradiente se estimaron hasta 500 y 1000 árboles respectivamente, según como mejoraba tras cada árbol la métrica de AUC en los datos de validación, donde las rondas de interrupción temprana fueron de 10. De forma similar mediante el uso del paquete \cite{RXgboost} se estimaron dos árboles de clasificación con potenciación de gradiente estimando hasta 500 árboles, según como mejoraba tras cada árbol la métrica de AUC y AUCPR en los datos de validación, donde las rondas de interrupción temprana fueron de 30. Una vez los cinco modelos se elige el modelo con mayor AUC y mayor área bajo la curva resultante entre las métricas precisión y exhaustividad resultante de la variación del umbral de probabilidad.

Una vez se haya elegido algún modelo se realiza un análisis de variables importantes. En este análisis se busca entender mejor qué variables considera más relevantes el modelo y como los distintos valores de estas variables afectan la predicción. Para lo primero, en cada modelo podemos medir distintas métricas, por ejemplo en el modelo logit es posible analizar los valores de los $\beta$ resultantes, los modelos basados en árboles es posible observar el Gini resultante tras cada partición, el resultado de la función de ganancia, la frecuencia o uso y el número de observaciones de cada variable. Para lo segundo es posible usar los gráficos de dependencia parcial, estos permiten observar el efecto marginal en la predicción de un modelo ya entrenado. A partir de los gráficos podemos observar si la relación entre $\hat{Y}\_{jt}$ y alguna variable continua es lineal, monótona o más compleja. De forma similar también podemos comparar si cada modelo responde forma similar o distinta. La función de dependencia parcial representa la predicción promedio si son forzados toda la muestra a tomar un valor de una variable sin manipular el resto de valores de las demás variables. Sin embargo en el proceso existen desventajas, la primer es asumir independencia entre las variables predictoras y que pueden existir efectos heterogéneos ocultos al promediar las predicciones para cada valor de la variable \citeA{molnar2018interpretable}. En el modelo se usó el paquete estadístico \citeA{DALEX}.