TFM KSCHOOL: PREDICCIÓN DE AUDIENCIAS

Master Data Science ed18

JAVIER CASADO ASGUINAGA

05/09/2010

El presente documento contiene la memoria del trabajo final de master necesario para dar por finalizado el master en Data Science de KSchool

1. INTRODUCCIÓN

El objetivo de esta herramienta es ayudar a los planificadores de medios en su tarea diaria, esencialmente en el seguimiento de sus campañas publicitarias de televisión. La actividad en televisión se mide en base al GRP(gross rating point), que siginifica contactos o audiencia a la que se llega con cada anuncio de la campaña. El principal método de compra de televisión en España es el coste por GRP,por lo que las marcas pagarán dependiendo de cuantas personas vean su spot. Por esta razón, la estimación de la audiencia de cada canal es muy importante, ya que dependiendo de su precisión la campaña estará alineada o no con el presupuesto objetivo, previamente definido por el cliente.

Los planificadores de medios tienen que hacer un seguimiento diario de la campaña para ver cuan lejos del objetivo está la campaña para ajustarla cada día comprando o cancelando spots. Este proceso lleva mucho tiempo y tener una estimación más realista les ayudaría mucho optimizando los tiempos, controlando la campaña y siendo más precisos.

El método de medición se basa en audímetros, que miden la cantidad de personas que ven la televisión y controlan todos los canales minuto a minuto. Para los diferentes spots emitidos en un mismo minuto, la audiencia será la misma, y por eso la frecuencia de observación ha sido analizada cada minuto por cada canal como un registro. Kantar Media es la empresa oficial encargada de medir las audiencias, y funciona como la herramienta oficial para todo el mercado español, para canales, marcas y agencias.

Hoy en día los planificadores de medios hacen el seguimiento de las campañas teniendo en cuenta las audiencias que los canales les envían (documento con todos los pases y audiencias llamado adjudicaciones) para ponderarlas con estimaciones aproximadas que reflejan la sobre o subestimación que los canales suelen hacer, basadas en su conocimiento personal. Esto hace que el proceso dependa de la experiencia del planificador de medios y de las estimaciones de los canales, por lo que los planificadores tienen un trabajo arduo para ajustar la campaña.

A fín de reeducir al máximo la diferencia entre audiencia estimada y la real(teniendo en cuenta la dependencia de los factores mencionados anteriormente) este proyecto ha buscado un sistema de previsión basado en algoritmos de machine learning.

Este sistema de predicción se ha alimentado con datos históricos de tres años. Dado que la variable a preddicir es continua, se ha trabajado con modelos de regresión.

En este documento se encontrará una guía detallada de las fases del proyecto, las transformaciones realizadas y líneas futuras de desarrollo.

2. OBTENCIÓN DE LOS DATOS

Para este proyecto se han cargado los registros de las audiencias históricas para entrenar al modelo. Para ser más precisos, el modelo se ha cargado con la audiencia minuto a minuto de los diecisiete canales de mayor audiencia de 2017 a 2020. Los archivos cargados se han dividido por semanas, debido a la limitación en la cantidad de entradas que soporta la herramienta oficial desde la que se han exportado los datos (InstarAnalytics de Kantar Media). La frecuencia de las entradas se estudia por minutos.

Como ya se ha mencionado, la medición de las audiencias se realiza mediante un dispositivo llamado audímetro, propiedad de Kantar Media. La muestra de población se equilibra según la población del país, evitando cualquier sesgo. Los datos obtenidos se extrapolan a la población total.

Kantar Media es una empresa multinacional y privada que posee tanto los datos como la herramienta de explotación(InstarAnalytics). La mayoría de las agencias y canales contratan esta herramienta. Para el control de la metodología y la métrica de todo el proceso de medición, se ha establecido un comité. Este comité está formado por personas pertenecientes a agencias, canales y marcas, los tres players del mercado. Los datos utilizados han sido proporcionados por una agencia internacional de medios de comunicación.

3. DESCRIPCIÓN DE LOS DATOS UTILIZADOS

En esta sección se describirán las variables contenidas en los archivos de datos utilizados para desarrollar el modelo. Sin embargo, los datos originales tienen más variables que las descritas en este documento, pero esta información no era relevante para el objetivo de este proyecto. El conjunto final de datos tiene más variables que el procedente de la fuente original y también la información descrita en esta sección. Se trata de variables procedentes de diferentes conversiones realizadas para alcanzar los diferentes requisitos. Esto se especificará en las siguientes secciones cuando lleguen las fases del proyecto.

A continuación se describen las variables utilizadas:

- Título:contiene la hora y el minuto de emisión. Este es uno de los datos más relevantes para el proyecto debido a que la audiencia dependerá del tiempo de emisión. El rango de este campo va desde las 2:30 hasta las 26:29 minutos. La razón detrás de esto será cubierta más adelante.
- Título/Descripción:muestra el nombre del programa de televisión que se está emitiendo en ese minuto. Se ha usado como una variable de control pero se usará como una variable de pronóstico en el futuro.
- Cadena:esta variable es realmente importante ya que, como el tiempo, el canal es un factor crítico en términos de audiencia, dependiendo del canal se podría esperar un nivel de audiencia diferente. Esto ocurre no sólo por la programación de la televisión, sino también por los hábitos de los consumidores.
- Fecha: otro campo relevante ya que nos da información sobre todos los factores temporales como el día de la semana, el número de la semana o el mes en el que se emite la entrada. Como se verá más adelante en este documento, la audiencia presenta una estacionalidad completamente clara a lo largo del año y el patrón se repite según el mes, el día de la semana y el horario.
- Género: esta variable incluye el género de la observación. Los géneros contemplados por la herramienta (InstarAnalytics) son: programas de ventas, ciencia ficción, entretenimiento, revista, publicidad, concurso, música, cultura, deportes, religión, toros, artes escénicas y otros.
- Públicoobjetivo: retrata el objetivo hacia el que se orienta un programa. Los diferentes objetivos que la herramienta considera son: familiares, adultos, niños, desconocidos, no disponibles y otros.
- Productora: refleja al productor del programa de televisión que se está emitiendo en este momento. Esto es variable a lo largo de los años.
- Calificación de edad: esta variable revela la edad mínima recomendada para ver el programa de televisión que se emite en este minuto.

- Ind. 16+ (c/inv.): esta es la variable para predecir y contiene el porcentaje de la audiencia promedio que está viendo este programa de televisión / canal en este minuto en este día.
- Ind. 16+ (c/inv.).1: muestra la audiencia promedio en miles de personas. Se usa como una variable de control para ver que la suma de este campo para todos los canales (en un minuto específico) nunca será 0 ya que la audiencia total de televisión en un minuto específico nunca podría ser 0 personas.

Traducción realizada con la versión gratuita del traductor www.DeepL.com/Translator

En este apartado se describirán las variables contenidas en los ficheros de datos, que han sido utilizadas para el desarrollo del modelo. No obstante, los datos originales contienen más variables de las que aquí se describen, pero estas no son relevantes para lo que se pretende lograr. El dataset final también contiene más variables de las que aparecen en este punto. Son variables surgidas de distintas transformaciones, debido a distintas necesidades del proyecto. Éstas se explicaran más adelante, en el apartado de Fases del proyecto.

A continuación se describen las variables utilizadas:

- **Título:** contiene la hora y minuto de emisión. Este dato es uno de los más relevantes para el proyecto, ya que la audiencia depende en gran medida de la hora de emisión. El rango de este campo va de las 2:30 hasta las 26:29. Más adelante se explicará esto en detalle.

- **Título/Descripción:** contiene el nombre del programa que se está emitiendo. Se ha utilizado como variable de control y en un futuro se pretende utilizar como predictor.
- **Cadena:** esta variable también es de una gran importancia, ya que al igual que la hora, la cadena marca en gran medida la audiencia que vas a tener. Esto se debe no solo a la programación, si no a los hábitos de consumo del espectador.
- **Fecha:** otro campo muy relevante ya que aporta información sobre le día de la semana y el mes en que se encuentra la observación. Como se verá más adelante, la curva de audiencia tiene una estacionalidad, y su patrón se repite en función de hora, día y mes.
- **Género:** contiene el género donde se enmarca la emisión de la observación. Los géneros contemplados son programas de ventas, ficción, entretenimiento, información, publicidad, consursos, múscia, culturales, continuidad, deportes, religiosos, toros, artes escénicas y otros.
- **Público objetivo:** esta variable plasma el público al que va dirigido el programa de la observación. El público contemplado es familiar, adultos, niños, desconocido, no disponible y otros.
- **Productora:** contiene la productora del programa que se está emitiendo. Variable a lo largo de los años.
- **Calificación de edad:** parecido al público objetivo, esta variable contiene la edad considerada como mínima para ver el programa.
- **Ind. 16+ (c/inv.):** esta es la variable a predecir, y contiene la audiencia media en porcentaje.
- **Ind. 16+ (c/inv.).1:** esta variable será utilizada solo para el control de datos, ya que no se pueden sumar porcentajes, y expresa la audiencia media en miles.

4. FASES DEL PROYECTO

El proyecto se ha dividido en tres fases diferenciadas, que se han ido desarrollando en paralelo a medida que se avanzaba. Cada una de estas fases se corresponde con un documento del repositorio. En este apartado se explican las tres fases y los pasos seguidos en cada una de ellas.

a. LIMPIEZA Y TRANSFORMACIÓN DE DATOS

Esta fase ha sido la que más tiempo ha necesitado. Se ha ido modificando los datos a la par que se avanzaba en la exploración del mismo. Se debe tener en cuenta para las transformaciones que todos los datos extraidos son del tipo string.

El primer problema que se encontró fue que la columna Título, la que contiene la hora de emisión, tenía intercalados los títulos del bloque de programación. Para eliminarlos de la columna hacemos un filtrado en el que solo nos quedamos con los valores que empiecen por '<<', ya que solo los datos horarios tienen esta característica.

A continuación, se observó que muchos campos del data frame, no solo la marca horaria, venían con los símbolos '<< >>' delante y detrás del registro contenido. Para eliminarlo se recurrió a un bucle que recorre la tabla entera y va eliminando los símbolos y almacenando las columnas modificadas en un nuevo data frame vacío.

El siguiente paso fue eliminar todos los registros que no perteneciesen a cadenas de los grupos comerciales de Atresmedia, Mediset o Pulsa, ya que estos tres grupos concentran el 93% de la audiencia total.

Tras ese filtro, se procedió a dividir la columna que contenía la marca horaria en hora y minuto. Esto se hace porque el rango horario va desde las 2:30 hasta las 26:29, ya que la herramienta de donde se extraen los datos no trabaja el rango horario convencional.

Debido al problema de rango horario, la fecha también presenta ciertas anomalías. En concreto se observa que los registros que van desde las 24:00 hasta las 26:29 contienen la fecha del día anterior, teniendo que sumarles un día para corregir el problema. Por ello se transforma la fecha a tipo datetime, para así poder sumar un día a los registros correspondientes.

Para completar la transformación anterior, se debe convertir las columnas hora y minuto al tipo entero. De esta manera, a tarvés de un simple flistrado, se obtienen los índices de aquellos registros cuyo valor de hora sea mayor de 23. Con esos índices se accede a los registros y se le suma 1 día a la fecha.

En este punto se crea la columna Machine_hour a partir de la columna hora. Esta columna se crea por los requerimientos de la fase de visualización. Se crea una columna nueva en vez de quedarse con la columna hora para así poder realizar transformaciones sobre la columna original.

En la columna hora se sustituye el 24 por un 0, el 25 por un 1 y el 26 por un 2. De esta manera la marca horaria queda en rango convencional, y la fecha queda ajustada a su valor real.

A continuación se vuelva a cambiar el tipo de las columnas fecha, hora y minuto. Se retoma el tipo original, string, para así poder unir estas tres columnas en una sola que compondrá la fecha completa. Esa columna completa se convierte a tipo datetime para poder realizar distintas operaciones con ella.

Después se procede a eliminar la columna Fecha y Título. Recordar que la columna Título era la que contenía originalmente la marca horaria.

El siguiente paso fue sustituir la coma por el punto en las columnas de audiencia. Esto se hace para poder convertir los valores a float. Una vez sustituida la coma se pasan las columnas de audiencia al tipo float.

Tras este paso, se procede a una de las transformaciones más críticas del proceso. Se observa que hay múltiples registres para un mismo valor de cadena y de fecha completa. Estos registros repetidos no comparten programación ni audiencia. Esto es debido a que hay cadenas que emiten en nacional y regional. Tras analizar profundamente esta casuística se observa que el valor máximo dde la audiencia de estos valores repetidos es igual a la suma del resto de audiencias. Por ello se hace una agrupación por fecha completa y cadena y se escoge el valor máximo de audiencia.

Como uno de los últimos pasos a seguir en esta fase se crean una serie de columnas a partir de la fecha completa. Estas columnas en su mayoría serán utilizadas en el apartado de visualización, y una pequeña parte será utilizada para el modelo. Las columnas creadas son Year, Date, Day_name,Month_name, Day_number, Month_number y Week_number.

Para generar las últimas columnas se crean 5 funciones distintas. Estas funciones se encargan de agrupar cadenas en grupos comerciales (Atresmedia, Mediaset y Pulsa), agrupar cadenas por canales de emisión simultánea (Multi, Grupo Cuatro y NSF), agrupar registros en base a la franja horaria en que se emiten (Morning, lunch,after, prime time, late prime time y night) y una última función encargada de generar el número ordinal del día de la semana.

Como último paso antes de guardar el data frame se renombran las columnas con la etiqueta final que se desea usar

b. MODELO

Dado que el objetivo del proyecto es predecir una variable continua, se utilizarán modelos de regresión para esta fase. A continuación se mencionan los cuatro modelos contemplados para obtener la predicción:

- El primer modelo valorado es el SARIMAX. Este modelo, específico para series temporales, tiene unos altos niveles de exigencia a nivel de dato, así como garn dicficultad en la implementación y optimización de hiperparámetros. Otro de los problemas encontrados es que con este algoritmo no se pueden realizar predicciones para cadenas distintas con un mismo modelo. Esto se debe a que es requisito indispensable tener un índice único del tipo timestamp. A persar de haber realizado un par de pruebas con este modelo, fue el primero en descartarse por los motivos expuestos.
- El segundo modelo valorado es el RandomForestRegressor. Este modelo se engloba dentro del aprendizaje supervisado, y funciona juntando varios árboles de decisión que no interactúan entre sí. Para sacar la estimación final se realiza de la media de todos los árboles. Este modelo es menos exigente con los datos que el anterior, y su implementación es relativamente sencilla, pudiendo obtener resultados sin necesidad de profundizar tanto el funcionamiento interno del modelo.
- El tercer modelo valorado es el GradientBoostingRegressor. Este modelo es similar al anterior, solo que los arboles de decisión que lo componen se van

ejecutando secuencialmente. De esta manera se busca que cada arbol de decisión corrija los errores del anterior. Los árboles de decisión que componen este modelo no suelen se muy profundos. Habitualmente rondan una profundiad de 2 o 3.

- El cuarto modelo es el LightGBMRegressor. Este modelo es muy parecido al anterior, solo que en vez de crecer horizontalmente como pasa con el GradientBoostingRegressor, lo hace verticalmente. Esto facilita que se reduzcan pérdidas en comparación con modelos anteriores.

Para comparar los modelos entre sí, inicialmente se utilizó el error medio absoluto. De esta manera se observó que el LightGBMRegressor era el que mejores predicciones daba. Una vez seleccionado el modelo, se utilizó también el error cuadrático medio para comprar modelos con distintas configuraciones de hiperparámetros.

Para la optimización de los hiperparámetros se intentó utilizar GridSearchCV, pero debido al gran tamaño del archivo, entorno a los 30 millones de registros, no se llegó a obtener resultados.

Finalmente se utilizó RandomizeGridSearch con listas de hiperparámetros reducidas. Una vez se hallaba el modelo óptimo, se eliminaban los hiperparámetros no óptimos de la lista y se añadían parámetros nuevos de orden mayor.

c. VISUALIZACIÓN

Esta parte se explica en detalle en el manual presente en el repositorio.

5. LÍNEAS FUTURAS

En este apartado se proponen una serie de puntos a desarrollar en un futuro. La idea es implementar modificaciones y procesos determinados para optimizar el modelo y poder hacerlo más certero y automatizado, así como implementar nuevas funcionalidades comerciales que puedan darle mayor salida a la herraienta. Las mejoras planteadas van desde modificaciones en la codificcación de las variables categóricas, hasta la generación de contenido relevante para el cliente.

A continuación se explican punto por punto las mejoras técnicas, a nivel de modelo, sugeridas para desarrollos futuros:

- En el modelo actual, la columna que contiene el título del programa de emisión no es tenida en cuenta. Una de las mejoras planteadas pasa por conectar la columna de título de programa con una base de datos que contenga un rating de los programas. A través de la nota obtenida en el rating, se codificaría esta variable, agregándole más peso a los registros que más nota tengan.
- Otra posible mejora, es añadir columnas basadas en fechas críticas que supongan grandes aumentos en las audiencias. Estas fechas pueden ser festivos nacionales, elecciones generales, eventos deportivos o similares.
- Debido a ciertas limitaciones de hardware, al tratarse de un desarrollo privado sin el respaldo y los recursos de una compañía, se ha procesado el proyecto a pequeña escala. La idea es escalar este proyecto al resto de cadenas e incluso otros mercados con los que realizar comparaciones, como Italia o Portugal. También se tendrían en cuenta el resto de segmentaciones de audiencia, pudiendo hacer una planificación más adoc para cada uno de los targets

comerciales. Para manejar ese volumen de datos, se propone hacer uso de los servicios cloud de pago, bien sea de Google o de Amazon.

- Actualmente, el modelo no es capaz de captar con precisión los registros correspondientes a publicidad ni las bajadas de audiencia que suceden. Esto es debido a que la variable género, que es la que marca, entre otras cosas, si un registro es publicidad o no, no atribuye el género publicidad a todos los registros que lo son. Esta situación mejoraría creando una función capaz de localizar esos registros y modificar su género a publicidad.
- Como última mejora técnica, se propone desarrollar un proceso que permita realizar predicciones sobre las adjudicaciones de la cadena. Esto cambia completamente el formato de dato utilizado hasta ahora, ya que las agrupaciones son distintas para cada cadena.

También es importante tener en cuenta las mejoras comerciales. Éstas dotarán a la herramienta de funcionalidades añadidas que le darán una mayor salida comercial. A continuación se describen esas mejoras:

- Implementar una función que permita optimizar la parrilla televisiva en función de la audiencia histórica del catálogo de emisiones.
- Complementar el archivo de visualización con gráficas de las tendencias del mercado por cadena, que permita tener un mayor poder de negociación con las mismas.
- Generar newsletters para el cliente con evolutivos del mercado, así como contenido para redes sociales que den mayor visibilidad a la compañía en el sector.