Tipología y ciclo de vida de los datos: Práctica 1

Samuel Campo Martínez y Marc Valdivieso Merino

14 de abril de 2020

1 Introducción

¿Existen diferentes clases de evoluciones para un valor en bolsa? ¿y para una cuota de una apuesta? ¿Cómo afectan las noticias a los espectadores en la visión del resultado final de un partido?¿Se puede visualizar la esperanza del público? ¿Cómo juegan y cómo jugar con el dinero y el futuro? Estas son las que nos hemos hecho poniendo el foco en las "Casas de Apuestas". Con el fin de comprender su comportamiento necesitamos extraer los datos con los que generar información y conocimiento respondiendo a estas y otras preguntas.

2 Contexto

Queríamos información veraz sobre algún tema que se pudiera conseguir sin buscar los datos sobre un usuario concreto. La cuota de una apuesta representa la creencia que tienen las personas que apuestan en ese partido, con la certidumbre de que los usuarios tienen un incentivo por apostar por lo que realmente creen (la devolución de dinero en caso de ganar la apuesta), y presentada por la casa de apuestas de manera que no afecta en absoluto a la privacidad de sus usuarios. En este sentido, la casa de apuestas actúa como un filtro de privacidad completamente efectivo y muy informado. Hemos cogido muchas casas para poder hacer filtro de posible ruido (Cuotas que no estuvieran bien representadas por favoritismo de un equipo u otro en una de las casas).

Tras esa elección se buscó la mejor fuente de datos y encontramos que hltv era una plataforma genial donde se mostraba mucha información de muchos de los partidos que sucedían. Además existían las cuotas de diferentes Casas de Apuestas y se juntaba con la información oficial de cada partido: resultado, jugadores, racha, etc. Al no tener una API oficial tuvimos que decantarnos por hacer scraping al código fuente de las páginas.

A pesar de esta organización visualmente sencilla, hemos comprobado las dificultades del proceso de scrapping ya que, por ejemplo, teníamos que la relación entre resultados o jugadores y partido no era directa y se tenía que hacer pasos intermedios para crear esa conexión y componer un dataset completo.

Las ideas que nos habíamos planteado en la introducción se iban viendo posibles de responder a medida que estudiabamos la estructura de la web y la posibilidad del código.

3 Dataset

3.1 Título

CS:GO Bets Time series

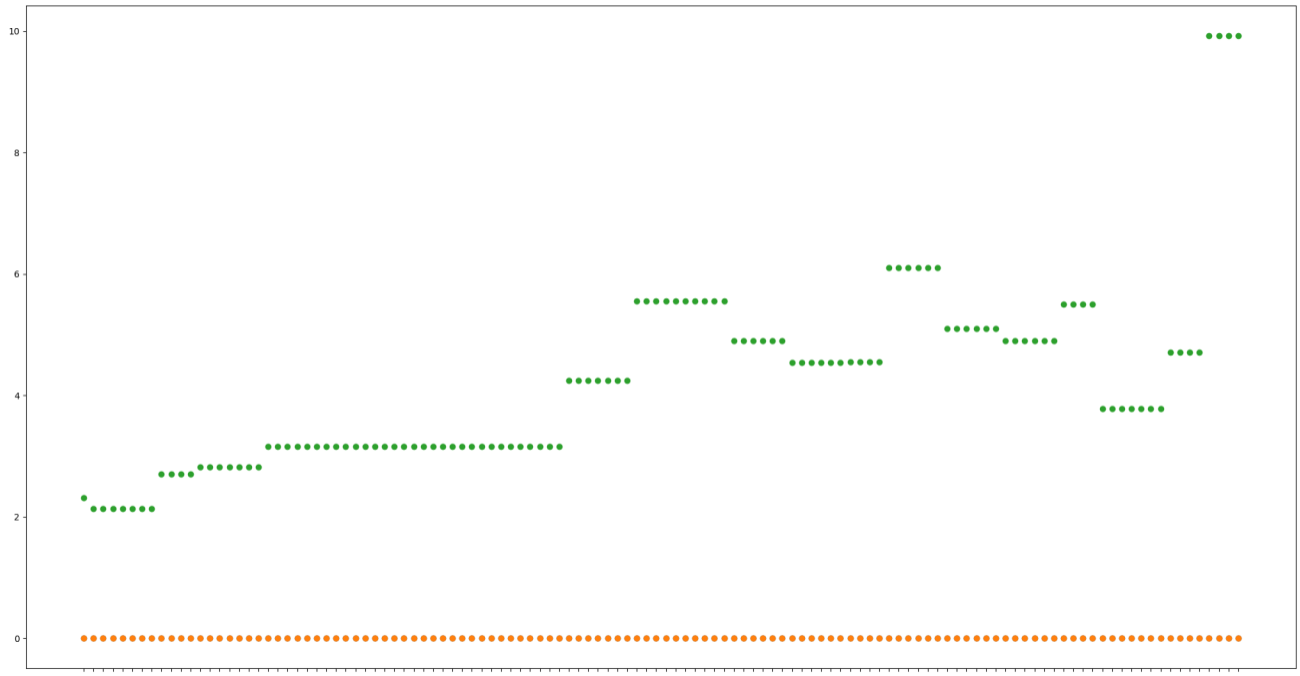
3.2 Descripción breve

Es un dataset con información sobre las cuotas de las apuestas de juegos del e-sport "Counter-Strike: Global Offensive" en las distintas casas de apuestas que admiten apuestas sobre éste, y con información capturada en tiempo real durante los partidos para ilustrar la fluctuación de la cuota durante estos.

3.3 Contenido

El dataset se organiza en 7 columnas: Bet, que es un float que representa la quota que da cada casa de apuestas (el factor de multiplicación de la cantidad apostada en caso de que la apuesta sea acertada). "Living", que dice si está o no en vivo el partido (variable categórica). Match, que contiene un ID de partido que relaciona la columna de equipo con el equipo rival (Un ID de match corresponderá a 2 equipos), Score, que dice la puntuación correspondiente al timestamp del equipo como entero, Team, que contiene el equipo, y timestamp, que contiene la fecha exacta del momento en que se ha capturado la cuota.

3.4 Representación gráfica



El gráfico muestra la evolución de las cuotas en función del tiempo contrastada con la información de mapas ganados. Una mayor esperanza de que un equipo gane corresponde a una cuota menor.

3.5 Inspiración

Tenemos dos intereses grandes de cara a capturar este tipo de datos: Por una parte, ver qué efecto tiene cada evento en tiempo real de un partido en la creencia de la posibilidad de victoria de los equipos en los usuarios, en un futuro pudiendo comparar cada uno de los eventos (En el caso de los e-sports, rondas ganadas, por ejemplo) a la fluctuación de la cuota media de las casas de apuestas. Además, sería muy interesante poder hacer scraping sobre webs de noticias deportivas (la idea original era hacer esto sobre fútbol, pero hay poca disponibilidad de partidos y no conocemos fuentes de este tipo de prensa en el caso de los e-sports) para intentar medir qué efecto tenía sobre las creencias de la probabilidad de ganar de cada uno de los equipos cada tipo de texto y en qué webs, y de tal manera intentar extraer información de qué tipo de comunicación con el público tiene más efecto.

Teniendo los datos actuales, podemos hacer un primer análisis sobre la fluctuación de la cuota en función de la puntuación en vivo.

3.6 Licencia

Tanto el código como los datos pueden ser usados por terceros pero teniendo que hacer referencia y dar crédito a nosotros como autores. Por tanto elegimos la licencia CC BY-SA 4.0 que permite seguir la traza de la autoría de los datos y del código, tal y como se explica en https://creativecommons.org/licenses/by-sa/4.0/. La cláusula sobre el uso comercial no nos interesa ya que probablemente perderíamos bastantes usuarios. Y la ODbL o la CC0 eran demasiado libres y ponían en peligro la autoría de nuestro producto.

3.7 Agradecimientos

Hemos scrapeado los datos desde https://www.hltv.org/ . Los datos son recogidas por ésta desde la serie de casa de apuestas indicada en su sección bets. También hemos usado datos de la propia HLTV, como los ID de los partidos que retransmiten, así como la presentación de equipos que tienen en su base de datos y los correspondientes datos asociados.

4 Metodología del scraping

4.1 Estructura de ficheros

Hemos definido un scraper para cada una de las url's de donde sacamos información: Uno para las cuotas en vivo de cada equipo para las que estén programadas desde cada una de las casas de apuestas (scraper-hltv-dds.py), uno para identificar qué partidos están en vivo (scraper-hltv-live.py). Y scripts adicionales que consiguen información de otras partes de la web para hacer relaciones entre los datos scrapeados y presentar un formato conreto del dataset.

4.2 Generación del dataset

Primero se inicializa el scraper-odds para conseguir al información de la tabla de https://www.hltv.org/betting/money . El mismo scraper reorganiza la información en un dataframe en cuatro columnas: Partido, cuota, proveedor de la quota, equipo y timestamp. Hay dos columnas adicionales: Live y Score. Si sabemos que un partido está en vivo, se indica en la columna pertinente y se añade la puntuación del partido en score.

Contribuciones	Firma
Investigación previa	SCM, MVM
Redacción de las respuestas	SCM, MVM
Desarrollo código	SCM, MVM