WIZELINE®

IA EN ACCIÓN:

Aplicando modelos de Machine Learning a la Copa Mundial de Fútbol 2018™

IA EN ACCIÓN:

Aplicando modelos de Machine Learning a la Copa Mundial de Fútbol 2018™

En el verano, la Copa Mundial de la FIFA™ 2018 fue un éxito rotundo. Los fans de fútbol soccer de todo el mundo se conectaron para ver el torneo en Rusia que tanto esperaban. Este caso de estudio profundiza en el experimento de Wizeline para predecir al ganador del torneo de la FIFA™ 2018 usando algoritmos de predicción.

¿El resultado? Una herramienta que superó al 95% de los participantes humanos. Compartiremos las probabilidades condicionales usadas por Wizeline y revisaremos los algoritmos de Machine Learning (Aprendizaje Automático) que las compañías pueden aplicar para resolver problemas de negocio únicos.

En sólo un corto periodo, las máquinas han probado comprender mejor los modelos de predicción que sus contrapartes humanos. Mientras algunas personas creen que son grandes visionarios, capaces de superar aún las mejores probabilidades, la mayoría confían en datos bien analizados por expertos antes de apostar.

Nos propusimos construir un algoritmo de predicción que pudiese competir con los mejores. La Copa Mundial de la FIFA™ 2018 presentaba la oportunidad perfecta para probar el algoritmo. También presentaba la oportunidad de fomentar competencia amistosa y cultura de equipo en Wizeline.

Nuestros científicos de datos aceptaron gustosamente el reto de crear una herramienta de predicción para estimar la probabilidad de que cada país avanzara, y finalmente ganara la Copa Mundial™. Lo llamamos Paul, por el famoso pulpo que "predijo" el resultado de la Copa Mundial de la FIFA™ 2010 en Sudáfrica. Sin embargo, nuestro Paul tiene variables cuidadosamente definidas, hipótesis científicas y opera basado en datos, no en suerte.

Primero lo Primero ¿Qué es IA?

Inteligencia Artificial (IA) es el concepto de que las máquinas son capaces de hacer tareas de una manera inteligente. **Machine Learning** es una aplicación de la IA y el área de IA de más rápido crecimiento. El rápido crecimiento de tecnologías de IA en los últimos 10 años ha sido en gran medida debido a los avances en Machine Learning. Consiste en construir algoritmos que aprendan de la experiencia y hagan predicciones acerca de datos. **La idea principal detrás del Machine Learning es que las máquinas puedan aprender por sí mismas si se les da acceso a los datos.**

En este caso, la IA es un programa que imita características humanas, como predecir el resultado de un partido. Toma la forma de algoritmos computarizados de toma de decisiones.

Estableciendo el Marco de Referencia

Nuestro Modelo

Primero, vimos los goles anotados y los resultados de los partidos de los países participantes en la Copa Mundial™ en un periodo de dos años. Nuestro modelo sólo usa datos de partidos no amistosos, porque creemos que son una mejor representación de las proezas de un equipo.

Intensidad de Goles y Resultados de Partidos

El primer paso para predecir el resultado de un partido es estimar el número esperado de goles que el equipo A anote contra el equipo B. Desafortunadamente, muchos países que se enfrentarían en la Copa Mundial no han jugado uno contra el otro recientemente. Hay relativamente pocos partidos entre selecciones nacionales en el fútbol, y difícilmente alguno

Partido amistoso - un juego de exhibición que no tiene impacto en la clasificación de un jugador o un equipo, o en el que el impacto es muy reducido. Comúnmente se le llama juego de práctica o de pretemporada.

Partido no amistoso - un partido oficial que tiene un impacto directo en la clasificación del jugador o del equipo en una liga o torneo.

entre países de diferentes confederaciones. Para sortear esto, vimos el resultado entre partidos recientes del Equipo A, enfocándonos en el número de goles que anotaron. Estos datos los triangulamos después al factorizar en la relativa defensa del equipo B con respecto a sus oponentes.

Considera un partido entre Alemania y Suiza. En los últimos dos años, Suiza concedió un promedio de 0.60 goles por partido. Alemania jugó contra Australia en la Copa Confederaciones, con un resultado de 3 a 2 a favor de Alemania.

Australia concedió un promedio de 1.12 goles por partido. Por lo tanto, basados en el número de goles que Alemania anotó contra Australia, esperamos que Alemania anote $3 \times \frac{.60}{1.12} = 1.61$ goles contra Suiza.

Esperamos que Alemania anote menos goles contra Suiza porque Suiza tiene una mejor defensa que Australia. Alemania jugó 20 partidos no amistosos en el periodo de dos años, así que **promediamos los resultados calculados sobre todos los oponentes.** Después de eso, hicimos un análisis similar para Suiza porque las intensidades de gol no son simétricas. Esto significa que el número esperado de goles que Alemania anotaría contra Suiza, no es el mismo que el número esperado de goles que Suiza anotaría contra Alemania.

Prediciendo el Desempeño de Alemania

Alemania (equipo A)

Suiza (equipo B)

Australia (equipo C)

Defensa Relativa de Suiza =

Promedio ⚽ concedidos por 🇨🇭

Promedio ⚽ concedidos por 🇦🇺

Número Esperado de Goles por 🇩🇪

Promedio ⚽ Contra 🇦🇺 × Defensa Relativa de 🇨🇭

Podemos hacer el modelo del número de goles que el equipo A anota contra el equipo B, usando la distribución de Poisson con la intensidad de goles $\lambda_{A,B}$ como su parámetro.

La distribución de Poisson es una distribución utilizada para expresar la probabilidad de que un número de eventos ocurran durante un periodo determinado de tiempo, dado el promedio de veces que el evento ocurre sobre ese periodo.

Para conocer el resultado de un partido, todo lo que necesitamos es la diferencia de goles, *Diff* = *X* - *Y*, donde *X* y *Y* son el número de goles anotados por los equipos A y B, respectivamente. Si *X* > *Y*, la diferencia es positiva y entonces muestra que el equipo A gana; Si *X* = *Y*, la diferencia es igual a cero y entonces muestra que el partido termina en un empate; si *X* < *Y*, la diferencia es negativa y entonces muestra que el equipo B gana.

Distribución de Poisson de los Resultados de Partidos

Diferencia de goles =

de goles del Equipo A(X) − # de goles del Equipo B(Y)

Si X > Y, gana

Si Y > X, gana

Si X = Y, es un empate

Resulta que, como tanto *X* y *Y* están distribuidos utilizando la distribución de Poisson, Diff sigue una distribución Skellam.

Una distribución Skellam - Es la discreta distribución de probabilidad de la diferencia X - Y de dos variables independientes al azar X y Y, cada una distribuida con la distribución de Poisson con valores respectivos esperados de Lambda y Mu.

La distribución de Skellam hace simple el computar la probabilidad de los eventos citados. Cuando los lazos no se permiten, como en el caso de la ronda final de la Copa Mundial™, necesitamos evaluar las probabilidades de ganar en tiempos regulares, tiempos extra (tratados como 30 minutos de partido independiente), y durante la etapa de penales.

Simulaciones Monte Carlo

La Copa Mundial™ se compone de la etapa de grupos, en la cual hay ocho mini torneos de todos contra todos. Después de eso, los ganadores y segundos lugares de cada mini torneo compiten en una ronda de eliminación llamada fase final. Corrimos cientos de simulaciones Monte Carlo para obtener diferentes escenarios, porque hay mucha incertidumbre en los resultados de la etapa de grupos.

Una simulación Monte Carlo, o simulación de probabilidad, es una técnica usada para entender el impacto del riesgo e incertidumbre en modelos de pronóstico.

- Primero, simulamos el resultado final de cada uno de los 48 partidos para saber cuáles países avanzarían a la siguiente ronda.
- Segundo, calculamos la probabilidad de que cada país avanzara en la fase final (cuartos de final, semifinales, final, ganador), usando la recursión.
- Finalmente, promediamos las probabilidades de todos los escenarios para obtener probabilidades incondicionales.

Este modelo nos permite actualizar las probabilidades con resultados de partidos actuales y en tiempo real a medida que estén disponibles. También usamos estos resultados de partidos para aumentar el conjunto de datos cuando se estima la intensidad de los goles. No hay necesidad de correr simulaciones cuando se sabe quiénes serán los competidores en la *ronda de 16*.

Algoritmos en Acción

Usamos probabilidad estadística para calcular la posibilidad de que cada país avanzara en el torneo y finalmente ganara la Copa Mundial™. Así es como los algoritmos de Machine Learning se pueden incorporar en nuestro marco de referencia para lograr el resultado.

El Machine Learning se usa frecuentemente para tareas de clasificación y regresión. La clasificación involucra predecir una respuesta cualitativa, lo cual toma los valores de una de las diferentes categorías de tipo K. La regresión se enfoca en predecir un resultado cuantitativo. Ciertamente podríamos utilizar la clasificación para predecir las probabilidades de los diferentes resultados de los partidos (llamémosle ganar, empatar y perder), pero sería difícil relacionarlos con los ganadores y segundos lugares de la Fase de Grupos. Esto es porque los ganadores se derivan del número de puntos ganados, la diferencia de goles y el número total de goles anotados. En su lugar, nuestro objetivo debería ser aprender A,B, el número esperado de goles que el equipo A anotará contra el equipo B.

Podemos lograr esto al usar características que contengan información de la estructura de un equipo, su renombre y desempeño reciente, e incluso factores económicos del país al que representan. Ejemplos de estas variables pueden ser:

- Edad promedio de los jugadores
- Número de jugadores compitiendo en la Liga de Campeones de Europa
- La confederación tanto del equipo en cuestión como la de su oponente
- Su clasificación en la lista de la FIFA™
- Posibilidades de ganar extraídas de probabilidades de corredores de apuestas
- PIB per cápita, normalizado por el promedio mundial

Estas características pueden recogerse de datos históricos en partidos oficiales, así como de estadísticas contextuales en el momento de los encuentros. Las variables numéricas pueden representarse como la diferencia entre los dos equipos, mientras que información como la confederación de cada país deben codificarse como variables separadas. Finalmente, hay que notar que al usar el número de goles que cada equipo anota como la variable de respuesta, cada partido suma dos observaciones diferentes, una por cada equipo.

Machine Learning en Movimiento

En estadística y Machine Learning nada es gratuito. Esto quiere decir que no existe un único algoritmo que funcione para todos los problemas, ni un método que controle a los demás sobre todos los conjuntos de datos posibles. Y esto es especialmente cierto en lo que se refiere al modelado predictivo.

Si bien no podemos afirmar que las máquinas de vectores de soporte funcionan mejor que un árbol de decisión, o viceversa, es importante probar los algoritmos que sean adecuados para el problema o tarea en cuestión. Un buen científico de datos debe probar múltiples algoritmos y usar un "conjunto de pruebas" de datos para evaluar el rendimiento y seleccionar al ganador.

Árboles de Regresión

Un algoritmo de Machine Learning que podemos emplear es el de árboles de regresión. Un árbol de regresión intenta encontrar la respuesta correcta preguntando tan pocas preguntas, cuya respuesta sea Sí o No, como sea posible. Cada pregunta debe reducir significativamente el resto de preguntas posibles. El conjunto de preguntas es seleccionado de tal manera que la variación entre todos los casos de prueba es mínima. Al momento de la predicción, respondemos todas las preguntas en función de las características del nuevo caso. **La predicción simplemente es el promedio de los valores de respuesta para los casos de prueba que siguieron la misma ruta en el árbol.**

Bosques Aleatorios

Si bien los árboles son simples y útiles para la interpretación, por lo general no son competitivos en términos de precisión de las predicciones. Para sortear lo anterior, podemos construir un bosque aleatorio, el cual es un método que produce múltiples árboles y luego combina esas predicciones para encontrar una solución, ya que no podemos cultivar diferentes árboles a partir de los mismos casos de prueba, y generalmente es costoso (y difícil) obtener más datos.

Bosque Aleatorio - modelo de predicción que produce múltiples árboles de regresión y combina las predicciones individuales para encontrar una solución.

En lugar de lo anterior, podemos tomar muestras repetidas, con reemplazos, del único conjunto de datos de entrenamiento y construir un árbol separado para cada muestra. En cualquier conjunto de datos, algunas observaciones pueden aparecer más de una vez, mientras que otras pueden no aparecer en absoluto. Tal como la diversidad de pensamiento beneficia a los humanos a alcanzar mejores consensos, los Bosques Aleatorios se benefician de tener árboles descorrelacionados. Una manera de lograr esto es restringir el algoritmo para que sólo considere un subconjunto de características por cada división (es decir, cada pregunta) en cada árbol.

Finalmente, el acuerdo se obtiene al promediar los valores predichos por todos los árboles del bosque.

Al construir Bosques Aleatorios es crucial afinar los siguientes hiperparámetros:

- **Características Máximas:** el tipo de preguntas que está permitido hacer en cada división
- **Profundidad Máxima:** el número total de preguntas consecutivas que está permitido hacer
- **Tamaño del bosque:** el número de árboles que deben ser cultivados o entrenados

Logramos esto probando diferentes configuraciones y comparando el error resultante en las observaciones Fuera de la bolsa. En específico, obtenemos una predicción por cada caso de entrenamiento utilizando los árboles que no incluyeron el caso de entrenamiento en el conjunto de datos de muestra (es decir, la bolsa) y luego escogemos la configuración que se desempeña mejor.

Aunque los Bosques Aleatorios son una poderosa herramienta de predicciones, otros algoritmos de aprendizaje, como las Redes Neuronales Artificiales, deberían ser consideradas también.

En última instancia, el desempeño del enfoque de Machine Learning debería medirse en función de puntos de referencia comunes, como las probabilidades de una casa de apuestas o incluso una simple regla de clasificación que seleccione al país con el rango más alto de FIFA™ como el ganador de un partido. Nosotros recomendamos empezar con un modelo simple y luego aplicar iterativamente métodos más sofisticados.

Poniendo a Paul a Trabajar

Los algoritmos predictivos solamente son efectivos según su valor o utilidad, por lo que pusimos a Paul a prueba. Enfrentamos a Paul contra 253 entusiastas y conocedores Wizeliners. Los desafiamos a vencer a Paul, y aceptaron el desafío.

Los empleados podían escoger a los equipos que creían que ganarían o empatarían en cada partido, registramos todo en una tabla de clasificación para ver cuáles Wizeliners hicieron las mejores predicciones en tiempo real.

Las Reglas del Juego

Cada predicción acertada otorgaba 10 puntos. Los empleados tenían además un "comodín" que les daba la oportunidad de adivinar el ganador de la Copa del Mundo, sin importar su posición en la clasificación o de sus predicciones de partidos anteriores. La opción de seleccionar el comodín estuvo disponible solamente hasta el inicio de la segunda ronda de la fase de grupos y otorgaba a los empleados 30 puntos extra si acertaban.

La predicción de Paul para cada partido se podía ver una vez que la votación se había cerrado y el partido estaba en curso.

Momentos destacados del torneo

• Francia vence a Croacia en la final del Mundial

VS

•Bélgica e Inglaterra son eliminados en las semifinales

VS VS

•Brasil es eliminado en los cuartos de final

VS

•España y Portugal son eliminados en la ronda de 16

• En un giro inesperado, Alemania quedó eliminada en la fase de grupos. La selección alemana era una de las favoritas en la competición, e incluso muchos analistas habían predicho que ganaría el torneo.

Las Selecciones Más Votadas Durante la Ronda de Comodines

1. Alemania : **27 %**
2. Brasil : **22.3 %**
3. España : **16.7 %**
4. Francia : **8.4 %**
5. México : **8.4 %**

Exactitud de las Tendencias de los Wizeliners

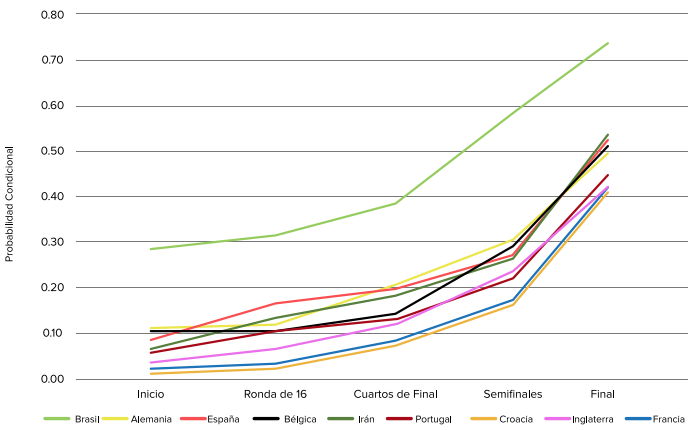
Calculado a partir de la participación de los empleados de Wizeline

- Los Wizeliners predijeron correctamente los resultados de 38 de los 64 partidos.
- Paul superó al 95 por ciento de los Wizeliners participantes.
- La Tendencia Wizeline, como predicción agregada, también superó al 95 por ciento de los participantes individuales

Información de los Gráficos de Probabilidad Condicional

Calculado antes del inicio del torneo y antes de que iniciara la fase eliminatoria

Probabilidad de Ganar, Condicionada a alcanzar cada etapa



Probabilidad de Ganar, Condicionada a alcanzar cada etapa

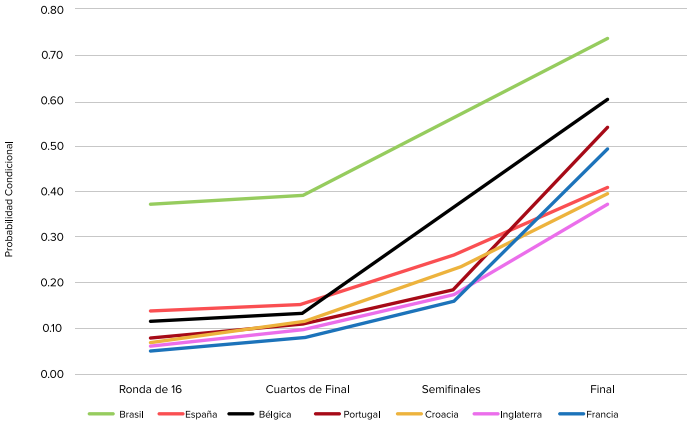


Gráfico #1: Irán fue un valor atípico en el modelo. De 10 partidos, Irán se mantuvo sin goles en contra en 9 de ellos, y sólo recibieron dos goles en el décimo partido. Por ello, Paul creyó que la defensa de Irán era fuerte. Esto es poco exacto cuando consideramos que Irán juega en una confederación poco competitiva.

Gráfico #2: En el segundo gráfico podemos ver que los países posicionados del lado izquierdo de la llave en el sorteo tenían más probabilidades de ganar el Mundial.

Usos Más Allá de la Copa del Mundo

Wizeline trabaja con clientes que se encuentran en distintas etapas de estrategias digitales o de su proceso de transformación. Queremos permitir que otras compañías aprovechen sus datos y los conviertan en información valiosa para sus negocios.

Entendemos la Inteligencia Artificial (IA) y podemos ayudar a nuestros clientes a construir la estrategia de IA correcta, utilizando el enfoque correcto, ya sea Machine Learning, probabilidad, ciencia de datos o crowdsourcing.

Más importante aún, no nos alejamos de los datos desordenados. Nuestros científicos de datos tienen experiencia construyendo aplicaciones prácticas de ciencia de datos, seleccionando un enfoque o modelo basado en los resultados comerciales esperados.

¿Cuándo deberían los equipos utilizar un enfoque determinado? ¿Cómo deben las empresas analizar las conversaciones provenientes de su audiencia? Wizeline ayuda a las empresas a descubrir cómo la ciencia de datos y el Machine Learning pueden agregar valor a sus negocios, reducir gastos y optimizar las áreas que mejor se desempeñan.