

Analytics' Workshop

Francisco Rosales Marticorena, PhD.

frosales@esan.edu.pe

11.12.18 – 12.12.18

ESAN Graduate School of Business

Tabla de Contenidos:

1 World Cup

2 Netflix

3 Big Data & Blockchain

World Cup

Classification / Recommendation

Financial services: price discrimination, i.e. insurance, loans.

Technology: user recommendations, e.g. Spotify, Netflix.

Actors in Peru

Corporations: BRECA data science challenge¹, Intercorp, etc.

Fintech: Bitinka, HolaAndy, etc.

Remark 1.1

Big data and blockchain are data management methods. Bitinka uses blockchain. HolaAndy uses big data.

¹Hackaton organized by Hackspace to recruit data scientists.

Popular Events

Google² → FIFA 2014 World Cup → 14/16 ($\approx 88\%$) accuracy

Goldman Sachs³ → Euro 2016 → got bookmakers' odds

Mister chip⁴ → Argentina–Peru → 53% of going to Russia 2018.

Google ⁵ says it can beat “Paul the Octopus”.

Beating “Paul the Octopus” is not trivial. Actually, it means getting an accuracy above 11/13 ($\approx 85\%$).

²<https://github.com/GoogleCloudPlatform/ipython-soccer-predictions>

³<http://www.goldmansachs.com/our-thinking/macro-economic-insights/euro-cup-2016/>

⁴From twitter account: @2010misterchip.

⁵Claim made by J. Tigani (Google I/O) at a big data conference (Strata) in 2014.

Objective of this Application

Replicate Google's idea:

- 1 Got the code → GitHub repository
- 2 Bought some data → OPTA⁶
- 3 Got the momentum → Peru vs. Argentina (as of 05.10.17)

Forecast Argentina–Peru:

- 1 Update data: Copa América Centenario & Qualifiers (so far).
- 2 Tweak parameters if necessary.

⁶www.optasportspro.com

Three leagues:

- 1 United States: Major League Soccer.
- 2 England: Premier League.
- 3 Spain: La Liga.

Features \times 2:

- | | |
|----------------------------------|--------------------------------|
| 1 Correct passes | 6 Shots |
| 2 Incorrect passes | 7 Corners |
| 3 Ratio correct/incorrect passes | 8 Cards |
| 4 Good passes at 80% top field | 9 Fouls |
| 5 Bad passes at 70% top field | 10 Expected goals ⁷ |

⁷Variable generated by OPTA.

Idea:

Make an assessment on how teams A and B are coming for the game based on their features⁸, and estimate the prob of A beating B.

Details:

- We keep only matches that resulted in a win/loss.
- We drop matches without complete information: new teams.
- We split the dataset in a test set (30%) and a training set (70%).

⁸features are computed in MA terms (last 6 games)

Statistical Model

- Consider the local-visitor model $y = f(x_1, \dots, x_m) + \epsilon$, where y_i is one if the local team won and 0 if it lost, x_1, \dots, x_m are explanatory variables for the outcome, and ϵ is some centered noise.
- For a collection of $i = 1, \dots, n$ observations and linear $f(\cdot)$ we obtain

$$y_i = \sum_{j=1}^m \beta_j x_{i,j} + \epsilon_i,$$

- Since y_i is binary, let $y_i \sim \text{Bernoulli}(p_i)$, i.e. $\mathbb{E}[y_i] = p_i$. Thus

$$p_i = \sum_{j=1}^m \beta_j x_{i,j} \rightarrow p_i = \phi \left(\sum_{j=1}^m \beta_j x_{i,j} \right); \quad \phi(z) = \frac{1}{1 + e^{-z}},$$

where $\phi(\cdot)$ is the logistic function, added to bound the RHS.

- Thus the model reads

$$p_i = \phi \left(\sum_{j=1}^m \beta_j x_{i,j} \right) \rightarrow \log \left(\frac{p_i}{1 - p_i} \right) = \sum_{j=1}^m \beta_j x_{i,j},$$

- We can use maximum likelihood to fit the model by maximizing the log of the joint probability

$$\begin{aligned} \mathcal{L}(y_i, p_i(\beta)) &= \log \left\{ \prod_{i=1}^n p_i(\beta)^{y_i} (1 - p_i(\beta))^{1-y_i} \right\} \\ &= \sum_{i=1}^n \log \{ p_i(\beta)^{y_i} (1 - p_i(\beta))^{1-y_i} \}, \end{aligned}$$

with respect to β .

- LASSO (what Google used), maximizes something very similar.

We aim to minimize

$$\mathcal{C}(\beta; \lambda) := -\log \left\{ \prod_{i=1}^n p_i(\beta)^{y_i} (1 - p_i(\beta))^{1-y_i} \right\} + \lambda \|\beta\|_1,$$

which is a **penalized** negative of the log-likelihood. Above $\|\cdot\|_1$ is the Manhattan norm, and λ is the regularization parameter.

Details:

- Regularization to avoid overfitting & selected via cross-validation.
- Automatic feature selection, i.e. some coefficients are zero.

Classification Accuracy:

The measure of accuracy of the model is given by

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP: true positive, TN: true negative, FP: false positive, and FN: false negative.

Results: 2014 FIFA World Cup

Replication of Google's algorithm

	Quarter Finals	Matches (no draws)	All Matches
Reported	88%	77%	–
Replicated	✓	✓	53%

Highlights:

- Google reported the quarter final's accuracy (only) at Strata, and the accuracy excluding draws (only) at its GitHub repository.
- The overall accuracy is not reported, but it is very relevant.

Results: 2018 FIFA World Cup (Quals)

Highlights:

- The accuracy excluding draws is 85%.
- The overall accuracy is 68%.

Details: Last 10 Games:

Team A	Team B	$\mathbb{P}(A > B)$	Expected	True
Uruguay	Argentina	66%	Uruguay	draw
Peru	Bolivia	72%	Peru	Peru
Ecuador	Brasil	7%	Brasil	Brasil
Paraguay	Chile	13%	Chile	Paraguay
Venezuela	Colombia	31%	Colombia	draw
Venezuela	Argentina	31%	Argentina	draw
Colombia	Brasil	46%	Brasil	draw
Chile	Bolivia	47%	Bolivia	Bolivia
Peru	Ecuador	58%	Peru	Peru
Uruguay	Paraguay	53%	Uruguay	Uruguay

Details: Coming 10 Games:

Team A	Team B	$\mathbb{P}(A > B)$	Expected
Peru	Argentina	39%	Argentina
Brasil	Bolivia	78%	Brasil
Ecuador	Chile	30%	Chile
Paraguay	Colombia	8%	Colombia
Venezuela	Uruguay	63%	Venezuela
Ecuador	Argentina	63%	Ecuador
Uruguay	Bolivia	63%	Uruguay
Chile	Brasil	37%	Brasil
Peru	Colombia	63%	Peru
Venezuela	Paraguay	37%	Paraguay

At 05.10.17 we did not know this results.

What do we need?:

- data
- python

Data and code are available at <https://github.com/arakata/FORECAST>

Netflix

- The Netflix prize (2009, 1M) → Belkor Pragmatic Chaos team.
- Examples: Amazon, Netflix, Pandora.
- Two types:
 - 1 Content-based filtering: look into the past.
 - 2 Collaborative filtering: look to others alike (we do this today).
- Optimization problem solved heuristically.

Optimization Problem

- Let $\mathbf{S} \in \mathbb{R}^{n \times m}$ represent the (sparse) rating matrix with users in the rows and movies in the columns. All we want to do is fill up \mathbf{S} .
- Consider the singular value decomposition

$$\mathbf{S} = \bar{\mathbf{U}}\bar{\Sigma}\bar{\mathbf{V}} = \mathbf{U}\mathbf{V},$$

where we can write $\hat{s}_{i,j} = \langle \mathbf{u}_i, \mathbf{v}_j \rangle$ for some i and some j , where \mathbf{u}_i is the i -th row of \mathbf{U} , \mathbf{v}_j is the j -th col of \mathbf{V} .

- The problem turns then into

$$\mathcal{L}(\mathbf{u}_i, \mathbf{v}_j) = \min_{\mathbf{u}_i, \mathbf{v}_j} \left\{ \sum_{i=1}^n \sum_{j=1}^m (s_{i,j} - \langle \mathbf{u}_i, \mathbf{v}_j \rangle)^2 \right\},$$

which can be solved by e.g. Stochastic Gradient Descent.

- ~ 10 thousand movies (m), ~ 600 users (u) and 18 movie genres (g).
- Let \mathbf{U} denote a matrix $10,000 \times 600$ of ratings, \mathbf{G} denote a matrix $10,000 \times 18$ of genres. Then

$$\mathbf{F} = \mathbf{G}^\top \mathbf{U}$$

is 18×600 , and the preferences of each of the 600 users are mapped in \mathbb{R}^{18} . Distance can then be measured some norm in that space.

MovieLens Example: R

```
> install.packages("recommenderlab")
> install.packages("ggplot2")
> install.packages("data.table")
> install.packages("reshape2")

> setwd("../AnalyticsWorkshop-master")
> source("movieRecF.R")

> recom_result
```

Big Data & Blockchain

- Blockchain is a distributed ledger with
 - 1 Pseudo anonymity
 - 2 Decentralization
 - 3 Perfect traceability
 - 4 Secure
- But is not only used to register “coins” of some kind.
 - 1 Augur
 - 2 Ujo
 - 3 OpenBazaar
- Machine learning technique can be potentially applied to this new data structure.

Disruptive Dapps: Arcade City

Uber in blockchain.

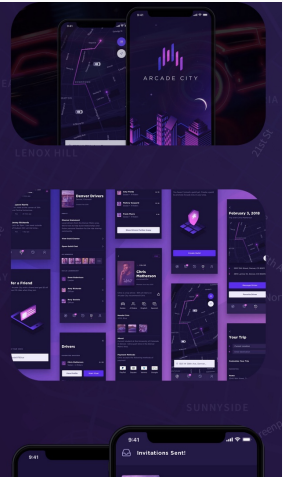


Source **Arcade City**

Disruptive Dapps: Arcade City

Uber in blockchain.




- **Driver Earnings** — Our top drivers earn 2-3x or more than they used to make with the corporate rideshares. And they do that all on their own terms, building recurring customer bases that no one can take away from them.
- **Driver Retention** — Uber and Lyft have between 50-90% annual driver turnover. That means **most** of today's Uber & Lyft drivers will **quit** in the next year. For Arcade City Austin that number is just **8%** — the lowest turnover and highest retention in the entire rideshare industry.
- **Cost Efficiency** — We've brought our operational expenses to maintain the Austin network down to **zero**. It is a truly independent, self-governing rideshare community operated by its drivers — and the very definition of a scalable model we are now excited to expand to New York and beyond.



Amazon in blockchain.

Sell Anything. Pay Zero Platform Fees.

Create a store. Sell whatever you'd like. Reach a new audience. Get paid in cryptocurrency.

 <p>Forever Tomorrow [EP] ★ 4.7 (192) \$4.99</p>	 <p>Summer Shades (original art) ★ 4.2 (23) \$500.00</p>	 <p>Blue Tank Top ★ 4.8 (242)</p>	✓ No Platform Fees
			✓ No Monthly Fees
			✓ No Listing Fees
			✓ No Bank / CC Required
			✓ Live Chat with Customers
			✓ Customize Your Store
			✓ Peer to Peer (no middleman)

Source **Openbazaar**

Bet365 in blockchain.



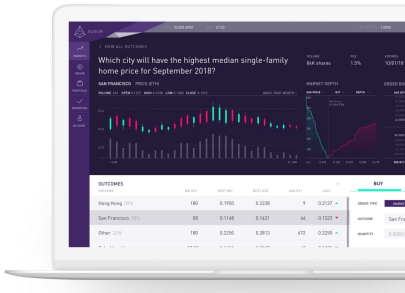
AUGUR

[Bug Bounties](#) [FAQ](#) [Blog](#)

The Future of Forecasting

A prediction market protocol owned and operated by the people that use it.

Get Started



Source **Augur**

Disruptive Dapps: Augur

Bet365 in blockchain.

The screenshot displays the Augur betting platform interface. At the top, the user's ETH balance is 0.1994, and their reputation (REP) is 0. The interface is divided into a sidebar menu on the left and a main content area on the right.

Sidebar Menu:

- MARKETS
- CREATE
- PORTFOLIO
- REPORTING
- ACCOUNT
- EOS ETHEREUM...
- SPORTS
- ETHEREUM BL...
- SPORT
- POLITICS
- NFL
- SPACE
- POLITICS, EVE...
- WORLD CUP
- CRYPTOCURRE...
- YOUTUBE
- GOLD
- FINANCE
- AUGUR
- ESPORTS
- ETHEREUM ME...
- US OPEN TEN...
- GOLF
- ETHEREUM
- MEDCREDITS
- CRYPTO
- PREMIER LEAG...
- TECH
- POLITICS, EVE...
- BOUNTIES
- EARTH
- DISASTERS
- WEATHER
- CAREER
- FINANCIAL MA...

Main Content Area:

Top Section: Formula 1 F1. OPEN OPEN INTEREST. SEARCH

Category: SPORTS. **Tags:** SOCCER / BARCELONA

Bet Title: Will FC Barcelona win against Real Madrid the next time they face on 28 October 2018?

Progress Bar: 0% to 100%. Current position: 51.70% (indicated by a green triangle pointing down).

Table:

VOLUME	FEE	EXPIRES
84.4780 ETH	1.0100 %	Oct 29, 2018 2:00 AM (UTC -5)

Buttons: ☆, TRADE

Category: SPORTS. **Tags:** FOOTBALL / SOCCER

Bet Title: Who will win the 2018/19 English Premier League?

Table:

Arsenal	4.50%	▼	Chelsea	11.00%	▲	Liverpool	20.00%	▼	+ 4 MORE
---------	-------	---	---------	--------	---	-----------	--------	---	----------

Table:

VOLUME	FEE	EXPIRES
0.3924 ETH	1.1211 %	May 13, 2019 2:00 AM (UTC -5)

Buttons: ☆, TRADE

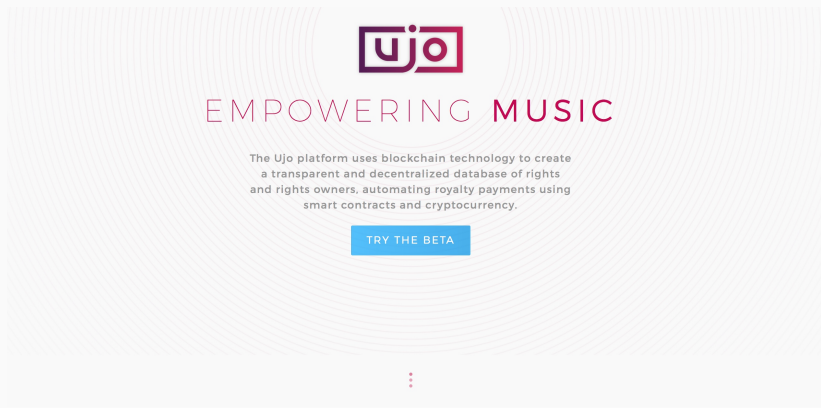
Category: SPORTS. **Tags:** FOOTBALL / SOCCER

Bet Title: Who will win the first 'El Clásico' match of 2018-2019?

Table:

Barcelona	50.00%	Real Madrid	50.00%
-----------	--------	-------------	--------

iTunes in blockchain.



Source **Ujo**

Disruptive Dapps: Ujo

iTunes in blockchain.


DISCOVER | MY COLLECTION

Ujo
BETA


REGISTER ARTIST

DISCOVER


Newest First




Funny Song
by: Bobby




Voyeur
by: Black Wolf




著作权法之网络快照
by: Chris




Mama Mia
by: Bobby




Facticity Devotion Discipline
by: Entropy Worship




Bunkerpop Theme
by: Bunkerpop



Super Mario
by: Nick Thelle



Pin Ball Kid
by: Dave Whiffin



Example 3.1 (El Problema de los Amigos)

Considere una economía formada por Usted y sus amigos: Alice, Bob y Charlie. Ustedes realizan actividades por las que se debe pagar con dinero, pero en ocasiones algunos de sus amigos no tienen efectivo, y otro debe cubrir el costo. Para llevar las cuentas claras deciden abrir una página web con un libro mayor para registrar las deudas y saldarlas luego.

Protocolo:

- Cualquiera puede agregar líneas en el libro mayor.
- Al final de cada mes todos se reúnen y pagan sus deudas en efectivo.

Problema:

Confianza. Si cualquiera puede agregar líneas en el libro mayor, es posible falsear deudas, e.g. Charly puede agregar una deuda (falsa) de Alice.

Definition 3.1 (Firma Digital)

Es una verificación digital que garantiza que su dueño ha visto la transacción y la autoriza. Para que un usuario tenga una firma digital, debe contar con una clave pública (CP) y una clave privada (CS).

Remark 3.1

La firma digital es “más segura” que la firma física en tanto cambia para distintos mensajes. En particular, considere:

$$\begin{aligned}\mathcal{F}(\text{Mensaje}, CS) &= \text{firma} \\ \mathcal{V}(\text{Mensaje}, \text{firma}, PS) &= \text{True/False},\end{aligned}$$

donde $\mathcal{F}(\cdot, \cdot)$ es la función que construye la firma digital, y $\mathcal{V}(\cdot, \cdot)$ es la función que verifica que la firma es legítima.

Remark 3.2

Aunque un ataque de fuerza bruta para encontrar firma tal que $V(\text{Mensaje}, \text{firma}, CP) = \text{True}$, es posible, es muy difícil, e.g. si firma tiene 256 bits, el ataque puede recorrer 2^{256} casos.

Remark 3.3

Note que si cada uno de sus amigos usa una firma digital, se puede estar seguro de que las deudas son legítimas. Sin embargo aún hay un problema, pues es posible copiar una misma línea varias veces. Este problema se resuelve agregando un contador a cada mensaje.

Protocolo:

- Cualquiera puede agregar líneas en el libro mayor.
- Al final de cada mes todos se reúnen y pagan sus deudas en efectivo.
- Sólo las transacciones firmadas son válidas.

Problema:

Confianza. Todos deben honrar sus deudas al final del mes.

Remark 3.4

Dado que sólo los deudores deben cancelar sus deudas en efectivo, es posible garantizar que no habrá deudores si ninguno de los amigos puede tener un saldo negativo (asumiendo que el libro mayor inicia con un saldo positivo $M_j > 0$ para cada amigo $j \in \{A, B, C, X\}$).

Protocolo:

- Cualquiera puede agregar líneas en el libro mayor.
- Sólo las transacciones firmadas son válidas.
- No saldos negativos.

Problema:

Confianza. El libro mayor es centralizado si e.g. está en una página web, i.e. es vulnerable.

Definition 3.2 (Libro Mayor Distribuído)

Es un libro mayor descentralizado que se actualiza en múltiples nodos de una red (e.g. los amigos) al mismo tiempo.

Problema:

¿Cómo saber que todos los nodos de la red están actualizando lo mismo?

Solución:

- Crear incentivos para validadores de libros mayores.
- Si hay dos libros mayores validados y distintos, confiar en el que ha sido más veces validado.

Definition 3.3 (Función Hash)

Una función hash es $f : \mathbb{X} \rightarrow \mathbb{Y}$, donde \mathbb{X} contiene cadenas de texto, o archivos, e \mathbb{Y} es una cadena de cierta longitud. Además pequeños cambios en $x \in \mathbb{X}$ generan grandes cambios en $f(x)$.

Example 3.2 (SHA256)

En este caso la imagen de f tiene 256 dígitos.

$$\text{SHA256}(\text{"arakata"}) \neq \text{SHA256}(\text{"Arakata"}) \neq \text{SHA256}(\text{"arakat4"}).$$

Definition 3.4 (Función Hash Criptográfica)

Es una función hash en la que la pre-imagen es muy difícil de calcular, i.e. la mejor opción es buscar en todo \mathbb{X} .

Definition 3.5 (Proof-of-Work)

Proof-of-Work es el trabajo por el que los validadores de los bloques son recompensados. El trabajo consiste en encontrar una cadena numérica (nonce), tal que

$$SHA256(LM, nonce) = \underbrace{0 \dots 0}_n \# \dots \# ,$$

n ceros

donde LM representa la información contenida en el libro mayor, y n es el número de ceros con el que empieza la imagen de la función hash. Si $n = 30$, se deben explorar 2^{30} casos. La prueba de trabajo es el nonce.

Remark 3.5

Los validadores de los bloques son llamados “mineros”. Actualmente su recompensa es de 12.5 BTC por bloque minado. Esta recompensa es el mecanismo por el cual la blockchain hace emisión monetaria.

De “libros mayores” a “bloques”:

- Considere el problema de los amigos.
- Suponga que al final del primer día se juntan todas las transacciones en un libro mayor (bloque): “Bloque 1”, y que a este bloque se le aplica una prueba de trabajo, de modo que es validado.
- Considere el mismo problema el siguiente día, pero adicionando en el encabezado del nuevo bloque “Bloque 2”, el hash correspondiente al último bloque validado, i.e. el “Bloque 1”, etc.

Remark 3.6

El conjunto bloques encadenados se denomina Blockchain. Es el libro mayor con una estructura que permite actualizarlo por bloques.