

Analytics con Python para Fintechs

Jessica Barbosa
Paco Mekler

14 de mayo de 2018

ITam

Agenda

- ¿Por qué Python?
- Análisis de datos productificado
 - Scraping
 - Limpieza
 - Feature engineering
 - Modelado y visualización
 - Diseño de producto
- Conclusiones

Código para esta sesión

- <https://github.com/mekler/Analytics-con-Python-para-Fintechs>
 - bases
 - scraper
 - limpieza
 - modelado
- Python 3.6.0
- Ipython notebook

Pequeña historia de Python

- La primera versión de Python es la 0.9.0 de 1991
- Existen dos versiones activas:
 - Python 2 (2.7.+)
 - Python 3 (3.6.+)
- ¿Por qué?
 - Hay muchas aplicaciones desarrolladas con Python 2 (soporte legacy)
 - Python 3 es mejor que Python 2 pero no se justifican los costos de migrar las aplicaciones

¿Por qué Python?

- Open Source
- Versátil
 - Desarrollo web
 - Batch Processing
 - Análisis
- Multi plataforma
 - Windows + Linux + Mac
 - Móviles (incluye IoT)
- Multi paradigma
 - Orientado a objetos, funcional, etc.

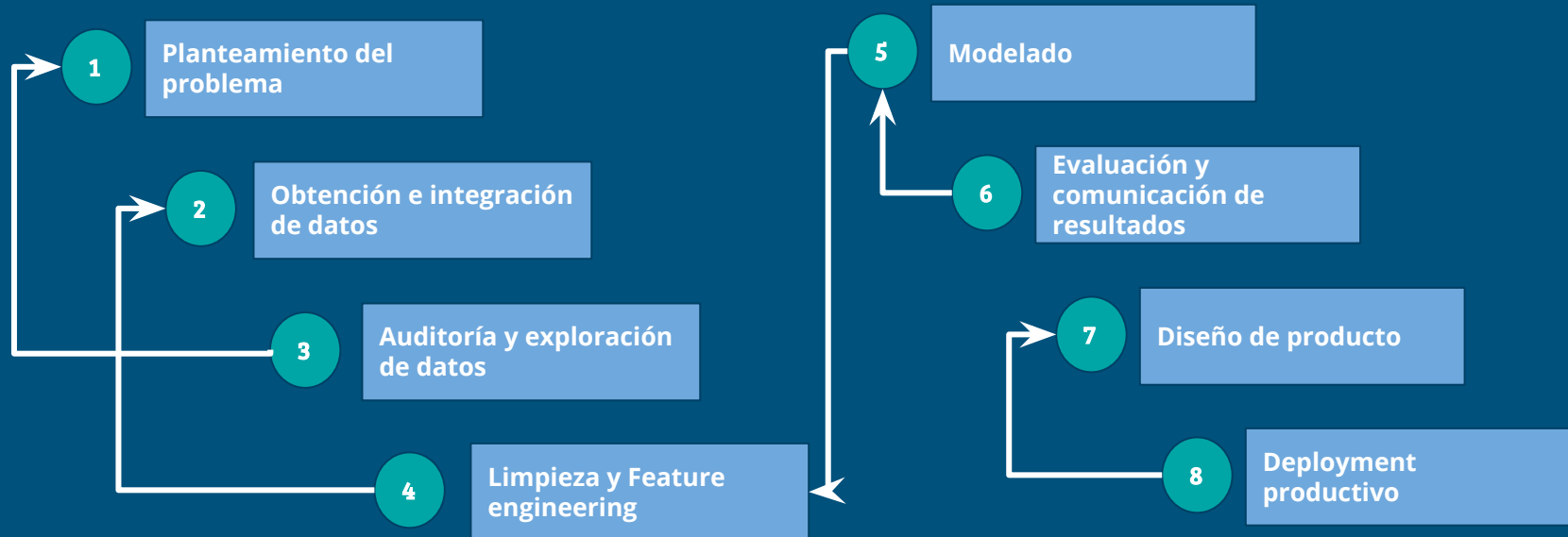
¿Por qué Python? (Ya en serio)

- No hay licencias => costos más bajos de implementación
- Un mismo lenguaje compartido por muchos equipos (i. e., desarrollo de software y análisis)
- Disminuye el problema típico “en mi computadora sí funciona” ... pero es linux
- Según mi experiencia con el lenguaje y mis necesidades, puedo hacer más cosas y de mejor calidad

¿Por qué Python para Fintech?

- Resumen ejecutivo: herramientas científicas
 - NumPy
 - SciPy
 - Matplotlib
 - Pandas
 - PyTables

Análisis de datos productificado



Código para esta sesión

- <https://goo.gl/MqSXQz>
- github.com/mekler/Analytics-con-Python-para-Fintechs

2. Obtención de datos: Scraping

- Extracción de información de fuentes no estructuradas
- Diferentes formatos
 - PDFs
 - XMLs
 - Páginas web

4a. Limpieza

- Normalización de información
 - Formatos de fechas
 - Catálogos
- Tratamiento de registros incorrectos
 - Falta de información
 - Errores en la información
- Conviene automatizarla si se hace frecuentemente
- Importante conservar los datos originales
 - Diferentes problemas pueden requerir diferente limpieza

4b. Feature Engineering

- Creación de indicadores que podrían ser útiles para explicar un fenómeno
- Combina diferentes fuentes de información
- Requiere análisis exploratorio de datos
- Es un proceso iterativo
- Herramientas:
 - Manuales (Excel)
 - Scripts en lenguajes de programación
- Más iteraciones => intentar *scriptear* lo más posible

5. Modelado y visualización

- Explicación de un fenómeno mediante datos
- Diferentes objetivos:
 - Clasificar observaciones
 - Predecir valor de una variable numérica
- Diferentes algoritmos
 - Regresiones lineales
 - Árboles
 - Redes neuronales
- Proceso iterativo

7. Diseño de producto

- ¿Producto de datos?
 - Heatmap en un PDF
 - Sistema que dé acceso a los datos y los resultados del modelo
 - Correo que avise de anomalías automáticamente
- ¿Quién es mi usuario final?
- ¿Quién lo va a desarrollar?
- ¿Quién lo va a mantener?
- ¿En qué infraestructura va a vivir?

Contacto y referencias

- José Incera (Maestría en Tecnologías de la Información y Administración)
 - jincera@itam.mx
- Jessica Barbosa
 - j.barbosa@opianalytics.com
- Paco Mekler
 - f.mekler@opianalytics.com
- Repo de esta sesion: <https://github.com/mekler/Analytics-con-Python-para-Fintechs>
- Tweets extraidos con: <https://github.com/juanjcsr/twitstream>
- Sección de modelado basada en:
[https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/Predict_Bay_Area_Home_Pr
ice.ipynb](https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/Predict_Bay_Area_Home_Price.ipynb)