Entregable -Proyecto

Oferta laboral en Data Science

06/08/2020 al 31/08/2020

Proyecto Individual

Mónica Villasuso López

THE BRIDGE - Data Science Boot Camp

Visión General

El objetivo de este proyecto es afianzar y ejercitar, de manera individual, los conocimientos adquiridos en la primera parte del Boot-camp de Data Science, particularmente los relacionados al proceso de análisis de datos (EDA) siguiendo todas las etapas del mismo (desde la obtención de los datos hasta las conclusiones derivadas del análisis).

Durante el último año, uno de mis objetivos personales se ha centrado en como orientar mi futuro profesional de cara a las necesidades actuales, la tecnología y tomando en cuenta mi perfil e intereses. Gran cantidad de estudios recientes, ubican a las profesiones relacionadas con los datos entre las más demandadas por las organizaciones debido a la transformación digital masiva y el incremento exponencial de la información almacenada que no sirve de nada si no es procesada adecuada y oportunamente.

Dado esto, y luego de revisar cientos de ofertas y sitios de empleo, de haber terminado en Marzo 2020 un curso de una de las principales herramientas de BI (Tableau), de haberme documentado sobre los principios "ágiles" con los que actualmente se gestiona el desarrollo de software y de haber leído muchos artículos relacionados con la tecnología decido, en Mayo 2020, formarme de forma intensiva en un Boot Camp en Data Science..

Ahora, a medio camino me pregunto: ¿me acerca esto realmente al mercado laboral? ¿Cómo es realmente la oferta? ¿Qué perfiles son los más demandados para entrar en este mundo? ¿Cómo son los salarios? ¿Es más conveniente pensar formarme en el área de Gerencia de Proyectos? Son muchas inquietudes y, antes de entrar en este prometedor mundo, quiero saber con certeza que dirección tomar. Entonces, ¿tiene seguido incursionar en "Data Science" sin haber analizado los datos de posibilidades de empleo en este sector? …sería un buen ejercicio para comenzar….

Objetivo

De acuerdo a los criterios de evaluación establecidos, he optado por el entregable A.

NOTA: El detalle con observaciones de cada apartado de los entregables se encuentra en el directorio **/documentation/entregables.xls** *También incorporado como anexo al final de este documento para facilitar su revisión*)

Especificaciones

Software

Python v2.7 o superior

Hardware

Procesador i5

Memoria RAM 8 GB

Espacio en disco 2 GB (para el dataset y los gráficos)

Requerimientos

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Os / Sys
- Json
- Plotly (*)
- Flask (*)

(*) Deben ser instalados previo a la ejecución de import de las librerías

Etapas

I. Contexto (Investigación)

- Luego de revisar los perfiles y roles relacionados con Data Science, los he agrupado en los 5 más comunes (detallados posteriormente).
- Para la revisión de los principales sitios de búsqueda de empleo consideré los más importantes actualmente: Indeed, Glassdoor y Linkedin. También tomé en cuenta la disponibilidad de datos y factibilidad/facilidad de obtenerlos.
- Debido a las restricciones de tiempo, seleccioné Glassdoor (el segundo sitio más grande) como fuente de información pues ya existía un data set disponible con el cual realizar el análisis.
- El data set contiene información de la oferta tecnológica en Glassdoor al 15 de noviembre de 2019.

- En el proceso de Data wrangling y Data cleaning se "mapeó" la oferta disponible a los roles definidos según lo siguiente:

  Data Analyst: Es el rol más básico. Se encarga de recolectar, procesar y analizar y reportar los datos desde el punto de vista estadístico.

  Business Analyst: Es un analista de Datos con un conocimiento del negocio tal que puede luego de analizar y comprender las necesidades y, a través del pensamiento crítico, agregar valor a la empresa sugiriendo nuevos análisis, definiendo casos de negocio y validando soluciones propuestas. Es un excelente comunicador que, además de gran dominio de herramientas de visualización e inteligencia de negocios (BI), también tiene un perfil técnico que le permite ser un enlace efectivo con el área de Tecnología de Información.

  Data Engineer: Es un perfil técnico, asociado principalmente al desempeño de los ambientes de datos desde el punto de vista de su arquitectura, entonación, seguridad, escalabilidad y rendimiento. Aquí se engloban los roles arquitectos de Big Data, administradores y especialistas en base de datos.

  Data Scientist: Es un perfil integral que cuenta con conocimientos estadísticos y matemáticos y además domina el software estadístico, los lenguajes de programación y los sistemas de análisis masivo y predictivo de datos. Todo lo anterior incluye además la estrategia y mecanismo de limpieza y depuración para garantizar su consistencia y la capacidad de comunicar resultados al negocio a través de las herramientas de visualización.

  Machine Learning / Artificial Intelligence: Es el perfil más especializado y avanzado de todo este grupo. Es experto en el aprendizaje automático a partir de los datos. Incluye también el Deep learning (aprendizaje automático para robots), la validación de modelos y el análisis de patrones de comportamiento de los datos. Es un rol cuyo aprovechamiento no está muy claro para muchas de las empresas aunque todas quisieran tener uno. Tiene un perfil matemático en combinación con un perfil técnico y se caracteriza por un continuo aprendizaje de las nuevas tecnologías.

  Project Manager: Perfil de gestión de proyectos caracterizado por su capacidad de organización, liderazgo y comunicación. Hábil para la toma de decisiones, identificación de conflictos y cuellos de botella, administración de recursos y con un conocimiento técnico y de negocio que le permita llevar a feliz término las iniciativas que involucran varias áreas de la empresa. Debe tener conocimiento de herramientas y metodologías que faciliten su rol.

  Other: Otros roles técnicos (principalmente desarrollo de Software y aplicaciones)

- Siendo los perfiles de análisis (datos y negocio) los más básicos y por ello, por los que es más fácil comenzar a desarrollarme en esta área, me pregunto si concentran una buena cantidad de la oferta laboral o si, por el contrario, el mercado está requiriendo en su mayoría perfiles más

avanzados o especializados con lo cual es más difícil tener acceso a ellos sin experiencia previa. Esto derivó en la siguiente hipótesis:

_____

"Al menos la mitad, (50%), de la oferta laboral en el área de Ciencia de Datos, corresponde a los perfiles de Análisis (Datos y Negocio)"

II. Obtención de los datos

Los datos fueron obtenidos del dataset en formato CSV de la siguiente fuente

https://www.kaggle.com/andresionek/data-jobs-listings-glassdoor

➔ Apartado 2 **/src/main.ipynb**

III. Data Wrangling / Data Mining / Data Cleaning

➔ Apartado 3 **/src/main.ipynb**

Creación de un subconjunto del Dataset original para lo cual fue necesario:

- Eliminar las columnas con información no relevante para el análisis
- Transformar los tipos de datos útiles para el análisis
- Estandarizar los códigos de países para poder utilizarlos posteriormente
- Eliminar los registros duplicados
- Estandarización de la información correspondiente al país
- Asignar experiencia requerida (Senior, Junior o Internship), nivel del puesto y tipo de trabajo (los roles explicados anteriormente) según la información contenida en el campo JobTitle.
- Completar con el valor NA (Not available) los campos en los que la información no exista (valores Nan,None, nulls)

RESULTADO DATA WRANGLING:

- De las 165.290 filas del DF original se eliminaron (por inconsistencias) alrededor del 35%, quedando 104.486 registros
- Se agregaron 6 columnas con información estandarizada y que facilita la interpretación
- Se redujo la memoria utilizada por el dataset en más de 90%% (De 193 MB a 19 MB)

IV. Exploración/ Análisis de datos (EDA)

El análisis de los datos está comprendido en el ***apartado 4 del jupiter notebook*** del proyecto (***/src/main.ipynb***). Los gráficos resultantes están almacenados en ***/resources/plots*** en formato html (para los gráficos dinámicos) o png (los gráficos estáticos). En particular, el análisis exploratorio de los datos está subdividido en:

4.1 Análisis preliminar de tendencias. Validación de hipótesis.

Estudio de las tendencias de los datos a través de una serie de gráficos de torta (pie) y gráficos de barras (bar) vistos desde cada una de las principales dimensiones de los datos .

Validación de hipótesis → **(RolDS_pie.png, RolDS_SB*.html)**

Análisis de la oferta laboral por → (**RolDS_*.html**, **Top_*.html)**

Bar Charts

- Industria
- Sector
- Tipo de empresa
- Empresas (Top 40)
- Experiencia
- Nivel del perfil requerido
- Tamaño de empresa

4.2 Análisis de salario por perfil → (**SalDS*.html, SalRol_Hist.html, Sal_Hist_Band_B5.html, Sal_Hist_Bandas.html)**

Estudio de los salarios, bandas mínimas, máximas y promedios por rol , antigüedad de empresa utilizando histogramas,

4.3 Análisis de Outliers → (**SalRol_Box.html, SalOutl_Scat.html, SalOutl_SectorPie.html**)

Utilizando gráficos de caja Gráficos de caja (para outliers) y gráficos de dispersión.

4.4 Análisis de distribución de la oferta en el mundo → (**JobsxRol_MM.html)**

Utilización de mapas para visualizar la distribución por rol en las diferentes regiones

4.5 Análisis de correlación → **(*_corr.png)**

Matriz de correlación visualizada como mapa de calor de las principales variables . Totalizada y por cada rol.

Apartado 5: → **(/resources/json/*.json y /resources/csv/*.csv)**

Exportación a formatos json y CSV de los dataframes depurados y usados en el proyecto

Apartado 6: → **(/resources/plots/Horas*.html)**

Análisis de las horas dedicadas al proyecto subdividido en cada una de las etapas.

V. Conclusiones

INDUSTRIA

- Sólo el 70% de los registros clasificados como roles de Data Scientist tenían la industria tipificada. De estos, las primeras 5 (en las que se concentra el 45% de la oferta) se concentra en la industria tecnológica y de consultoría.
- Si vemos la gráfica por industria únicamente para el rol de Business Analyst, el sector de Banca e Inversión desplaza a Internet y se coloca entre los 5 primeros.

SECTOR

- Los 3 primeros sectores (Tecnología de Información, Servicios de negocio y Finanzas) concentran el 60% de la oferta. Esta proporción se mantiene para todos los roles de DS analizados.

TIPO DE EMPRESA

- El 50% de la oferta proviene de empresas de naturaleza privada y el 32% de caracter público.

EMPRESAS

- Amazon concentra la mayor cantidad de oferta de empleo en general, sin embargo para los roles de DS, el primer lugar lo ocupa la empresa *Hays* (consultora de selección de personal especializada). Otras dos grandes empresas de selección de personal (M*yitjobs* y *Gotfriends* (israelí)) están también en los primeros lugares.) Amazon y las empresas de consultoría estratégica (como BCG, McKinsey) son, según esto, quienes más se están desarrollando en esta área a nivel especializado.

NIVEL DE EXPERIENCIA

- Aunque sólo el 25% de las ofertas en Data Science especificaban el nivel de experiencia requerida para el cargo (19.6% Senior, 4.2% Junior y 1.2 recién titulados), en el 80% de los casos buscan perfiles senior. Esto es más notable para el caso de los roles más especializados (Data Scientist y Machine Learning) donde el 85% de las ofertas requiere profesionales nivel senior (versus el 70% en el caso de Data Analysts).
- 80% perfiles Senior, 16.2% perfiles junior y solo 3.8% (recién titulados o en prácticas)

TAMAÑO DE EMPRESA

- Alrededor de 1/3 de las ofertas está en empresas pequeñas (menos de 200 empleados), 1/3 en ofertas de empresas entre 200 y 10.000 empleados y el otro tercio restante en empresas grandes (más de 10.000 empleados). Esta proporción se mantiene a lo largo de todos los roles analizados]

SALARIOS

- En las gráficas se observa con claridad como varia el salario en función al rol, siendo los menos retribuidos los roles asociados a Analistas de Datos, luego Analistas de Negocios y aumentando a medida que el rol va requiriendo más experticia técnica o más nivel de especialización (Ingenieros de Datos, Científicos de Datos y Machine Learning)
- Los salarios más altos están en empresas de antigüedad menor (más innovadoras y menos tradicionales) y de menor tamaño (desde el punto de vista de número de empleados)
- Para el total de la oferta en el área de Data Science, la banda salarial se mueve entre los 50.000 dólares anuales a 90.000 dólares anuales. Ambas curvas se asemejan a una campana de gauss y están dispersos en torno a los valores de los picos de la curva (con mayor dispersión para el caso del tope superior de la banda salarial donde hay mayor desviación con respecto a la media])
- Si disminuimos la cantidad de rangos (bins) a 5, disminuye la eficiencia del análisis y el mismo ya no es tan representativo de la realidad, pues perdemos precisión en como se distribuyen realmente los datos. Pocos bins son ideales cuando tenemos menos de 50 observaciones.
- Por el contrario, un número de bins muy grande es más representativo pero también dificulta el análisis. Por ello, para esta cantidad de datos (más de 100), es razonable la propuesta de la herramienta de 15 bins
- Viendo los datos de salario medio (media entre el límite inferior y superior de la banda), observamos que hay mayor dispersión en los mismos (y son más altos pues en frecuencia están más desplazados a la derecha) cuando el rol es más especializado o complejo

OUTLIERS

- Las empresas outliers son en su mayoría empresas tecnológicas pequeñas (menos de 500 empleados) o del sector salud (biotecnología y farmacéuticas) (más de 5000 empleados)
- Las empresas outliers en términos de salarios pertenecen principalmente (más de la mitad) a los sectores de Tecnología de Información, de investigación (farmacéutica, salud, biotecnología) y Aeroespacio y Defensa

DS EN EL MUNDO

- Los principales países con oferta en los niveles más especializados de Data Science son: Estados Unidos, Canadá, Alemania, Francia, Reino Unido, India e Israel

- Para los niveles más básicos (Analistas), se agregan a esta lista Suiza, Bélgica, Holanda y Australia.
- Europa es el continente donde se concentra la mayor cantidad de ofertas en esta área

CORRELACIÓN

- Para el análisis de correlación utilizamos las únicas 3 variables numéricas (Salario medio, Tamaño de la empresa y Antigüedad de la empresa (por año de fundación).
- Viendo los datos de manera global no se observa correlación alguna entre las variables, sin embargo al analizarlos filtrando por rol vemos que para el caso de Ingenieros de datos o Ciencias de datos se puede ver una correlación negativa baja (un poco mayor a 0.5) entre el año de fundación de la empresa y el tamaño de la misma. Esto se puede interpretar como que empresas más nuevas (año de fundación es mayor)]) son más pequeñas (en términos de cantidad de empleados). Esto es bastante lógico pues son profesiones muy recientes que han impulsado la creación, en los últimos años de empresas dedicadas a este rubro.

OTROS:

Lo que cambiaría

Tomaría datos más actualizados. La decisión sobre la fuente de datos utilizada fué principalmente por el corto tiempo.

Lo aprendido

El tiempo a invertir en el proceso de limpieza y depuración de los datos superó con creces mis expectativas y eso restó tiempo al resto de las fases del proyecto. Lo consideraré en proyectos a futuro.

Luego de este proyecto estoy bastante más familiarizada con el funcionamiento de los gráficos (como objetos), las librerías para graficar y sus diferencias, sin embargo me hubiera gustado tener más tiempo para explotar los "features" y características de cada uno y utilizarlos de forma más avanzada.

Como uno de los objetivos adicionales me hubiera gustado analizar las ofertas por rol para ver las herramientas más pedidas en cada uno (por ejemplo, Python, Spark, SQL, Tableau, Power BI), Tensorflow y afinar así mi formación

Tiempo dedicado al proyecto.

El tiempo total dedicado al proyecto fue de 102 horas

Un 63% del tiempo fue dedicado a las tareas relacionadas con los datos y de esto casi la mitad (32 horas) se consumió estandarizando y limpiando los datos para su posterior utilización (i.e. el 31% del tiempo total del proyecto)

ANEXOS

ENTREGABLES - PROYECTO Oferta laboral en Data Science		
	DETALLE	OBSERVACIONES
C	Document all steps. Structure your code to keep it cleaned using good practices.	
	2.Collect the data. Try to do each call, it collects the last updated data.	**Apartado 2 (/src/main.ipynb) - Jupiter Notebook** Los datos utilizados corresponden a un data set disponible en Kaggle de Noviembre 2019. Por temas de disponibilidad de tiempo se usaron estos datos en lugar de buscarlos actualizados.
	3.Determine and explain if the data is cleaned. If not, then clean it.	**Apartado 3 (/src/main.ipynb) - Jupiter Notebook**
	4.Create an API that returns a Json with the logic explained. The flask server must beexecuted running the src/api/server.py file.	**/src/api/ server.py** **/src/utils/apis_tb.py**
	5.Show different tendencies for each column in your dataset.	**Apartado 4.1 (/src/main.ipynb)- Jupiter Notebook**
	6.Represent, in a pie chart, the time you needed for each point in the The projectsteps section.	**Apartado 6 (/src/main.ipynb) - Jupiter Notebook**
	7.Answer the questions:a.Was it possible to demonstrate the hypothesis? Why?b.What can you conclude about your data study?c.What would you change if you need to do another EDA project?d.What do you learn doing this project?	**Conclusión (/src/main.ipynb) - Jupiter Notebook**
B	Show the histogram of each column of your dataset with bins=5. How are the rangespainted?	**Apartado 4.2 (/src/main.ipynb) - Jupiter Notebook**
	2.Which are the columns with the highest correlation? Draw the correlation matrix.	**Apartado 4.5 (/src/main.ipynb) - Jupiter Notebook**
	3.Use Matplotlib functions to show all graphs. No with pandas directly	**Apartado 4 (/src/main.ipynb)- Jupiter Notebook**
A	Research to save each plot in local files.	**/src/utils/visualization_tb.py** Todos los gráficos generados están disponibles en **/resources/plots**
	2.Use distribute modules for each functionality. The jupyter notebooks must not haveany loop or functions. It only must have the initials imports and the call to necessaryfunctions.	**/src/utils**
	3.Apart from matplotlib, use seaborn to show the graphs.	**Apartado 4 (/src/main.ipynb)** Adicional a Seaborn se utilizaron gráficos de la librería Plotly (Plotly express y Plotly graphs objects)
	4.Answer the questions:a.Are there outliers or some rare data?b.What are the columns that have more repeated values?	**Apartado 4.3 (/src/main.ipynb) - Jupiter Notebook**
A +	Create a pull request for the entire project.2.How can you put your flask server with a public IP without Heroku? realizethat flask starts the server in a private net as default (localhost)3.How can you put your flask server with a public URL without Heroku?4.How can you put your flask server with a public IP with Heroku?5.How can you put your flask server with a public URL with Heroku?6.Are there more urls from where to collect your data?. Explain why. If yes,then collect it and merge it with your data.7.In order to practice OOP and engineering/architecture concepts incomputing, define all the functions inside classes and make the programfunctional using them. After that, use a program to create the class diagram.8.Using your own API url, use web scraping to get the json and show the data	**Apartado 5 (/src/main.ipynb - Jupiter Notebook** En este apartado se exportan en formato json y csv (a los directorios correspondientes en /resources) los dataframes resultantes de la depuración y utilizados para el análisis. Mi idea era utilizarlos para el punto 8 de este entregable pero no fué posible