ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH THĂM DÒ BỘ DỮ LIỆU KAGGLE
ML/DS SURVEY 2021

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Hữu Trường	18521564

TP. HỒ CHÍ MINH – 12/2021

1. GIỚI THIỆU:

Từ năm 2017, Kaggle tổ chức các cuộc khảo sát từ đầu tháng 1 đến đầu tháng 4 với người sử dụng. Thông tin thu thập được, được tạo thành một dataset và sử dụng để tổ chức một cuộc thi phân tích dữ liệu thường niên mỗi năm.

Đề tài thực hiện phân tích trên bộ dữ liệu khảo sát của năm 2021, được thu thập từ khoảng 9/1 đến 10/4. Thu được khoảng 25 973 mẫu dữ liệu từ 171 quốc gia.

Bộ câu hỏi khảo sát gồm 39 câu hỏi chính 9 câu hỏi phụ, câu hỏi thuộc dạng trắc nghiệm chọn một hoặc nhiều đáp án. Kaggle đã loại bỏ sẵn các trường hợp duplicated và spam (thời gian khảo sát dưới 2 phút.

Nhận thấy bộ câu hỏi có đề cập đến mức thu nhập hằng năm, em làm đề tài này nhằm mục đích phân tích điểm khác biệt giữa người trong ngành DS/ML ở những mức thu nhập hằng năm khác nhau. Vừa rèn luyện khả năng kể chuyện từ dữ liệu vừa nắm được thông tin và xu hướng của cộng đồng DS/ML để cập nhật bản thân.

Kết quả nắm được phần lớn các insight theo nhu cầu đồ án mà dataset thể hiện.

2. NỘI DUNG

2.1. Giới thiệu bộ dữ liệu:

Dataset gồm 369 cột chứa đáp án của 42 câu hỏi chính và 9 câu hỏi phụ từ bộ câu hỏi khảo sát. Có nhiều cột như vậy là tại vì cách trình bày của các câu hỏi nhiều đáp án.

Để thuận tiện cho việc mô tả dataset thì em đã tiền xử lý dataset trước, nhận thấy không có bất cứ giá trị bị khuyết hay sai loại dữ liệu, không đồng nhất về cách trình bày nào, em đã gom 369 cột thành khoảng 49 cột feature, và gom nhóm lại theo nội dung của chúng, thể hiện trong bảng codebook dưới đây.

STT	Tên feature	Nội dung thể hiện
	Thông tin cá nhân	
1	Age range	Độ tuổi
2	Gender	Giới tính
3	Country	Quốc gia
4	Education level	Trình độ học vấn hiện tại
	Thói quen sử dụng	
5	Computing platform	Nền tảng tính toán
6	Specialize hardware	Phần cứng đặc biệt
7	IDE’s	Các IDE sử dụng
8	Hosted notebook	Các hosted notebook sử dụng

9	Place to publicy app	Nơi họ public các ứng dụng, sản phẩm của mình lên
10	Course platform	Nền tảng khóa học về ML/DS mà họ tham gia
11	Media source	Nguồn tìm kiếm thông tin về ML/DS mà họ hay sử dụng
	Thôn tin công việc cá nhân và công ty đang làm việc	
12	Role	Vị trí hiện tại trong công ty
13	Important activities at work	Công việc chiếm phần lớn thời gian ở công ty
14	Coding age	Kinh nghiệm viết code tính theo năm
15	TPU's using time	Số lần sử dụng TPU
16	Industry	Ngành của công ty hiện tại đang làm việc
17	Size of company	Quy mô công ty (tính bằng tổng số lượng nhân viên của công ty)
18	Number of people respond for DS/ML at company	Số lượng người phụ trách mảng DS/ML tại công ty đang làm việc
19	How company incorporate ML/DS to bussiness	Cách công ty áp dụng DS/ML vào công việc của họ
20	How much money you/team spend on ML/DS for last 5 years	Lượng tiền cá nhân, hoặc đội ngũ trong công ty đổ vào các ứng dụng, sản phẩm hỗ trợ DS/ML trong vòng 5 năm qua
	Xu hướng công nghệ	
21	Programming language	Ngôn ngữ lập trình đang sử dụng
22	Programming language for newbie	Ngôn ngữ lập trình mà họ khuyến khích người mới học nên sử dụng
23	Data visualization library	Thư viện trực quan dữ liệu họ sử dụng
24	ML frameworks	Framework máy học mà họ sử dụng

25	ML algorithms	Thuật toán máy học mà họ sử dụng
26	Computer vision method	Nhóm các phương pháp thị giác máy tính họ sử dụng
27	NLP method	Nhóm các phương pháp xử lý ngôn ngữ tự nhiên họ sử dụng
28	Cloud computing platform	Nền tảng tính toán đám mây
29	Most enjoyable cloud computing platform	Nền tảng tính toán đám mây ưa thích nhất
30	Cloud computing platforms hope to become more familiar with in the next 2 years	Nền tảng tính toán đám mây mong muốn học trong vòng 2 năm tới
31	Cloud computing product	Sản phẩm về tính toán đám mây mà họ sử dụng
32	Cloud computing products hope to become more familiar with in the next 2 years	Sản phẩm về tính toán đám mây mà họ muốn học trong vòng 2 năm tới
33	Data storage products	Sản phẩm về lưu trữ data mà họ sử dụng
34	Managed machine l earning products	Sản phẩm giúp quản lý học máy mà họ sử dụng
35	Managed machine l earning products ho pe to become more familiar with in the next 2 years	Sản phẩm giúp quản lý học máy mà họ muốn học trong vòng 2 năm tới
36	Big data product	Sản phẩm về big data mà họ sử dụng
37	Most often using bi g data products	Sản phẩm về big data mà họ sử dụng thường xuyên nhất
38	Big data products h ope to become mor e familiar with in the next 2 years	Sản phẩm về big data mà họ muốn học trong vòng 2 năm tới

39	Business intelligence tools	BI tools mà họ sử dụng
40	Most often business intelligence tools	Bi tools mà họ sử dụng thường xuyên nhất
41	Business intelligence tools hope to become more familiar with in the next 2 years	Bi tools mà họ mong muốn học trong vòng 2 năm tới
42	Categories of automated machine learning tools	Loại tools tự động hóa quá trình học máy mà họ sử dụng
43	Categories of automated machine learning tools hope to become more familiar with in the next 2 years	Những tool tự động hóa quá trình học máy mà họ mong muốn học trong 2 năm tới
44	Specific automated machine learning tools	Tools tự động hóa học máy cụ thể mà họ đang sử dụng
45	Specific automated machine learning tools machine learning tools hope to become more familiar with in the next 2 years	Tools tự động hóa học máy cụ thể mà họ muốn học trong vòng 2 năm tới
46	Tools to help manage ML experiments	Tools giúp quản lý và cải thiện trải nghiệm học máy
47	Tools for managing ML experiments hope to become more familiar with in the next 2 years	Tools giúp quản lý và cải thiện trải nghiệm học máy muốn học trong vòng 2 năm tới
48	Primary tool to analyze data	Tool chính dùng để sử dụng phân tích dữ liệu

49	Yearly compensation	Tổng thu nhập hằng năm

2.2. Phương hướng EDA:

Để phân biệt sự khác nhau giữa các nhóm thu nhập, dataset được chia thành 4 dataset nhỏ hơn theo 4 mức thu nhập:

+ Thu nhập hằng năm bằng 0

+ Thu nhập hằng năm dưới 10000 đô

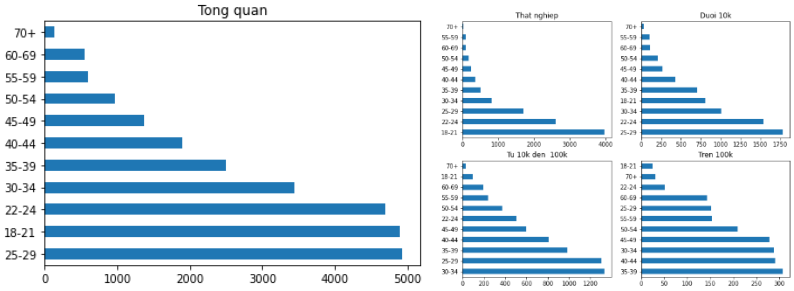
+ Thu nhập hằng năm từ 10000 đô tới 100000 đô

+ Thu nhập hằng năm trên 100000 đô

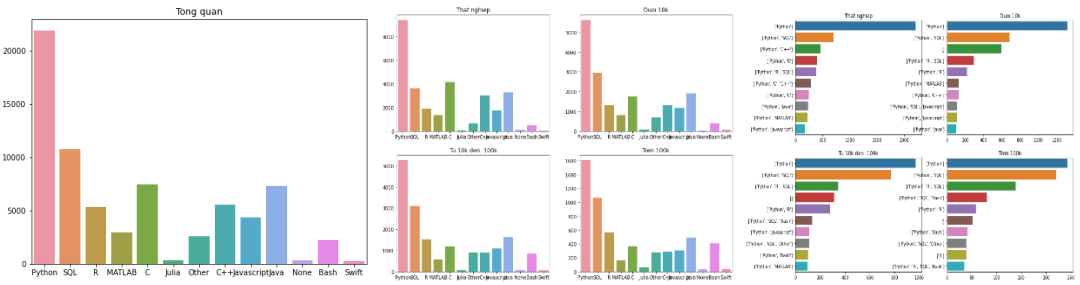
Việc phân tích thăm dò sẽ được thực hiện song song trên 4 nhóm này.

Có hai loại câu hỏi, câu hỏi một đáp án và câu hỏi nhiều đáp án. Mỗi loại câu hỏi có cách xử lý khác nhau và vì mọi feature của data đều giống nhau nên hai cách này sẽ được áp dụng xuyên suốt quá trình phân tích thăm dò trên toàn bộ dữ liệu.

Loại câu hỏi một đáp án, sẽ có thống kê tổng quan số lượng đáp án trên toàn bộ dataset và số lượng đáp án trên mỗi dataset con.



Loại câu hỏi nhiều đáp án, nhờ có thao tác gom cột ở bước tiền xử lý, không chỉ có thể thống kê số lượng đáp án mà còn thống kê được tổ hợp đáp án nào xuất hiện nhiều nhất trong mỗi nhóm thu nhập hằng năm.



2.3. Insight thu được:

2.3.1. Thông tin cá nhân:

+ Phần lớn người tham gia khảo sát từ 18 đến 30 tuổi, giới tính nam nhiều vượt trội, phần lớn đến từ ấn độ hoặc mỹ, đa số đều có ít nhất bằng thạc sĩ hoặc cử nhân.

+ Độ tuổi chưa có việc làm tập trung vào khoảng từ 18 đến 25 tuổi. Những người đã có việc làm đều từ 25 tuổi trở lên. Những người có thu nhập trên 10000 đô một năm từ 30 tuổi trở lên. Những người có thu nhập trên 100 000 đô một năm khoảng 35 tuổi trở lên. Rất ít người tham gia từ 55 tuổi trở lên.

+ Yếu tố giới tính không ảnh hưởng đến việc có công việc hay mức thu nhập.

+ Ở mỹ thì tỉ lệ có việc cao hơn tỉ lệ thất nghiệp nhiều, và những người trên 100000 đô một năm đều xuất phát từ mỹ. Riêng nhật bản là quốc gia có ít người tham gia nhưng số lượng người trên 30 tuổi lại nhiều hơn và đa số đều có việc làm.

+ Phần lớn người thất nghiệp có bằng cử nhân, mức lương càng cao, tỉ lệ số lượng thạc sĩ và tiến sĩ càng tăng lên.

2.3.2. Thói quen sử dụng:

+ Đa số sử dụng laptop hoặc máy tính để bàn, ở mức lương cao thì họ sử dụng thêm cloud computing platform.

+ Đa số đều không dùng phần cứng đặc biệt, nếu có thì NVDIA GPUS được ưa chuộng nhất.

+ Jupyter notebook và VS đối với offline , google colab và Kaggle notebook đối với online là được sử dụng phổ biến nhất.

+ Họ thường chia sẽ các sản phẩm của mình lên Kaggle, Github, Colab.

+ Họ rất chuộng các khóa học của coursea, udemy và Kaggle learn.

+ Đa số đều tìm kiếm thông tin trên youtube, Kaggle và các bài blog.

2.3.3. Thông tin nghề nghiệp cá nhân:

+ Đa số người tham gia là học sinh, và thường là chưa có việc làm. Công việc data sciencetist chiếm tỉ trọng nhiều nhất ở tất cả các mức lương. Ngoài ra thì dưới 10 000 đô một năm hoặc

từ 10000 đến 100000 đô một năm thường thấy ở data analyst và sofware engineer. Trên 100000 đô một năm có software engineer và research scientist chiếm tỉ trọng cao nhất.

+ Phần lớn người tham gia chỉ có từ 1 đến 3 năm kinh nghiệm coding. Nhóm người có dưới 1 năm kinh nghiệm coding thì phần lớn là thất nghiệp hoặc lương dưới 10000 đô một năm. Từ mức lương 10000 đến 100000 đô một năm thì ít nhất phải có 1 đến 3 năm kinh nghiệm và tốt nhất là trên 3 năm kinh nghiệm viết code. Mức lương trên 100k chiếm phần lớn là người từ 5 đến trên 20 năm kinh nghiệm viết code.

+ Công việc phổ biến nhất là phân tích và hiểu dữ liệu phục vụ cho mục đích của công ty ở mọi mức lương. Ngoài ra còn có việc xây dựng, sửa chửa, vận hành bảo trì các cơ sở hạ tầng dữ liệu. Lương càng cao càng thấy sự gia tăng của các nhiệm vụ: dựng prototype để explore applying machine learning to new area, cải thiện mô hình máy học hiện có, xây dựng ML servcie, nghiên cứu những công trình state of art của ML.

+ Đa số đều chưa sử dụng TPUs bao giờ.

2.3.4. Thông tin nghề nghiệp về công ty đang làm việc:

+ Chủ yếu là doanh nghiệp thuộc ngành Computer/Technology. Ở mức lương dưới 10 000 nhiều người trong ngành Academic/Education. Nhóm Academic/Education giảm đi khi lên mức lương cao hơn, thay vào đó là Accounting/Finace, Medical/Pharmaceutical.

+ Những công ty có quy mô khoảng 50 nhân viên trở xuống thường trả lương dưới 10000 đô một năm. Những công ty có quy mô khoảng 1000 nhân viên tới trên 10000 nhân viên thì mới trả trên 10000 đô một năm.

+ Với mỗi công ty hoặc thường là team nhỏ từ 1 đến 2 người (công ty nhỏ) phụ trách hoặc là một team lớn trên 20 người (công ty lớn) phụ trách mãng data science. Những công ty lương trên 10 000 đô đa số là tầm 20 người trở lên phụ trách mảng Data Science.

+ Đa số đều không chi trả cho các dịch vụ hỗ trợ. Khoảng 1000 đến 10000 cho 5 năm với mức lương mỗi cá nhân 10 000 đến 100 000. Hơn 100 000 đô một năm với những công ty, cá nhân có mức lương trên 100 000 đô một năm. Đối với mỗi cá nhân chi khoảng 0.2% tối thiểu cho các dịch vụ. Đối với tập thể team lớn, mức lương cao, chi trả nhiều hơn.

+ Đa số những người chịu trách nhiệm về mãng data science trong công ty đều trong giai đoạn đầu, chỉ mới nghiên cứu và chưa áp dụng Machine learning methods. Với mức lương dưới 10 000 hầu như không biết hoặc chỉ đang nghiên cứu thêm các ML methods. Với mức lương 10

000 đến 100 000 đô thì có những công ty chuẩn bị áp dụng ML method vào bussiness. Với mức lương 100 000 thì đa số đều đã áp dụng thành công ML methods hoặc mới bắt đầu áp dụng.

2.3.5. Xu hướng công nghệ sử dụng:

+ Python là ngôn ngữ được sử dụng nhiều nhất. Xu hướng ngôn ngữ đều là như nhau. Python, SQL, R, C, C++ và JAVA là những ngôn ngữ sử dụng nhiều. Bộ ba ngôn ngữ được sử dụng nhiều nhất là Python, R , SQL. Ở mức lương 10 000 trở xuống thất nghiệp thì thường thấy xuất hiện C, C++, Java, Javascript, MATLAB. Ở múc lương 10 000 trở lên có sự xuất hiện của Bash. Đa số đều khuyên những người mới học thì nên học python.

+ Data visualization library: matplotlib, seaborn, scikit-learn. Matplotlib phổ biến nhất.

+ ML frameworks: Đứng đầu là tensorflow, theo sau đó là keras, pytorch và xgboost.

+ ML algorithms: Đứng đầu là linear, logistic regression, decision tree, random forest. Ở mức thu nhập thấp thì thường dùng linear/ logistic regression và decision tree hoặc random forest. Ở mức thu nhập cao Gradient boosting machine (xgboost, lightgbm) được sử dụng nhiều hơn.

+ CV methods: Phổ biến nhất là các phương pháp Image classification. Tuy nhiên các tác vụ khác như object detection, image segmentation, general purpose image/video tools cũng có số lượng không ít.

+ NLP methods: Những phương pháp NLP được sử dụng phổ biến nhất là các phương pháp Word embedding (Glove, word2vec,..) và các mô hình ngôn ngữ Transformer (Bert, GPT-3, ..vv), các mô hình encoder-decoder (seq2sed,...).

+ Cloud computing platform: AWS là phổ biến nhất, ngoài ra còn có Google cloud platform và microsoft Azure. Tất cả đều cho rằng các nền tảng trải nghiệm như nhau.

+ Cloud computing products: Dưới 10000 đô thu nhập một năm phần nhiều là google cloud computing. Trên 10000 đô thu nhập một năm là EC2. (Amazone elastic compute cloud).

+ Data storage product: Amazone web storage service (S3) và Google cloud storage (GCS) là hai storage product thường được sử dụng nhất.

+ Managed ML product: Google Cloud Vetex AI phổ biến với mức lương thấp dưới 10 000 đô. Amazone sagemaker, data bricks, azure machine learning studio là ba product phổ biến nhất.

+ Big data product: MySQL, PostgreSQL and Microsoft SQL Server là ba big data product được ưa chuộng nhất. Trong đó MySQL được sử dụng nhiều vượt trội nhất.

+ Business intelligent product: Tableu và power BI.

+ Automated ML tools: Data augmentation, hyperparameter tuning và model selection là các tool tự động phổ biến. Với mức lương dưới 10000 đô một năm: thường sử dụng các tools model selection và data augmentation(nhiều nhất). Với mức lương 10000 đến 10000: hyperparameter tuning, model selection (nhiều nhất) và full auto ML pipeline. Với mức lương trên 10000: người ta ít sử dụng model selection chỉ còn hyperparameter tuning(nhiều nhất) và full ML pipeline

+ Tools help experience ML: TensorBoard hoặc MLflow

+ Tool to analyze data: R studio, Jupyter Lab hoặc những phần mềm cơ bản (Excel, Google Sheet,..).

3. KẾT LUẬN

Có thể tóm gọn các insight thu được từ quá trình phân tích thăm dò như sau.

Mức 25 tuổi và có bằng thạc sĩ trở lên sẽ giúp chắc chắn hơn về việc làm và mức lương. Mỹ là nơi có nhiều việc làm lương cao nổi trội hơn trung bình các quốc gia còn lại.

Các nguồn tham khảo thông tin và public bài phổ biến bao gồm Github, Kaggle, youtube, các khóa MOOC coursea, udemy. Tìm hiểu thêm về cloud computing platform như Amazone web service.

Có khoảng trên 5 năm kinh nghiệm coding và có thêm khả năng xử lý các công việc đa dạng sẽ mang lại thu nhập cao hơn. Công việc chính trong ngành vẫn là phân tích dữ liệu theo mục đích của công ty.

Phần nhiều các công ty, doanh nghiệp trong ngành Academic\Education thu nhập không cao bằng công ty Công nghệ hay Y tế, những công ty nhỏ khoảng dưới 50 nhân viên có đội ngũ DS ít và lương thường nhỏ hơn các công ty lớn trên 10000 nhân viên.

Phần lớn các xu hướng công nghệ phổ biến được đề cập đều là các xu hướng công nghệ được giảng dạy ở trường đại học. Dù ở mức lương nào thì xu hướng công nghệ vẫn như nhau.

Toàn bộ dữ liệu ở dạng categorical và hầu như không có quy luật nên không thể phát triển mô hình.

TÀI LIỆU THAM KHẢO

[1] Cuộc thi Kaggle survey 2021. Link: https://www.kaggle.com/c/kaggle-survey-2021

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	18521564	Lên ý tưởng, làm bài, viết báo cáo, thuyết trình.