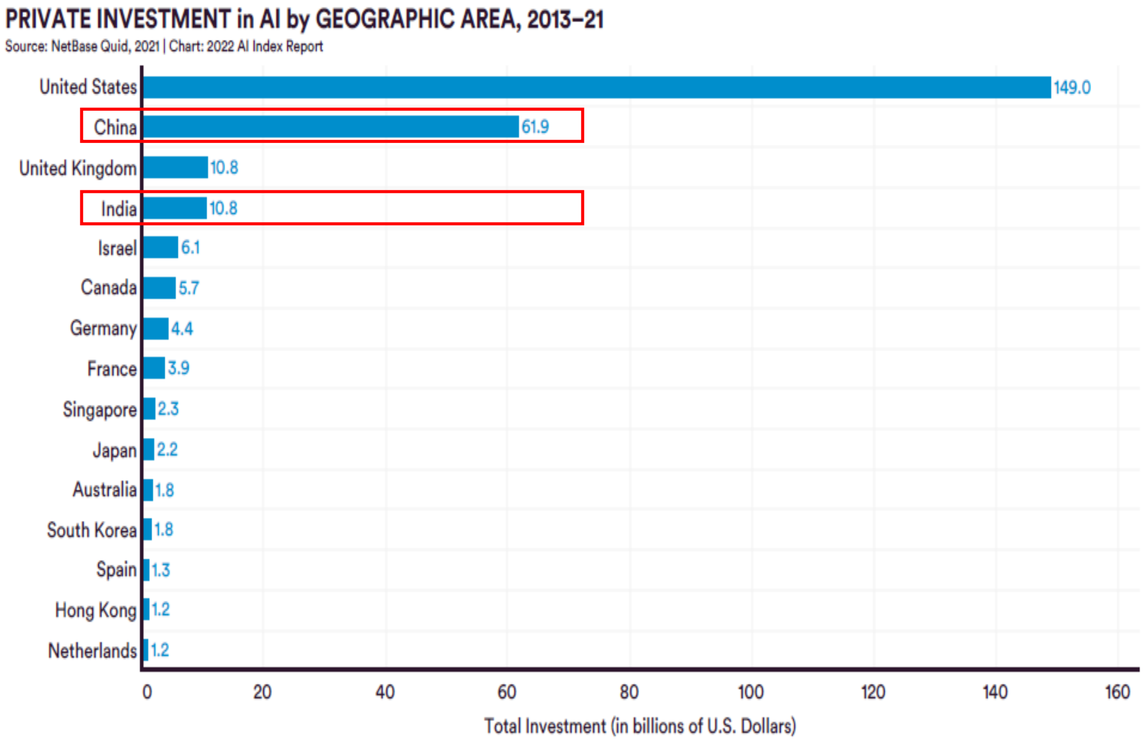
How can inexperienced Data Scientists increase their income?

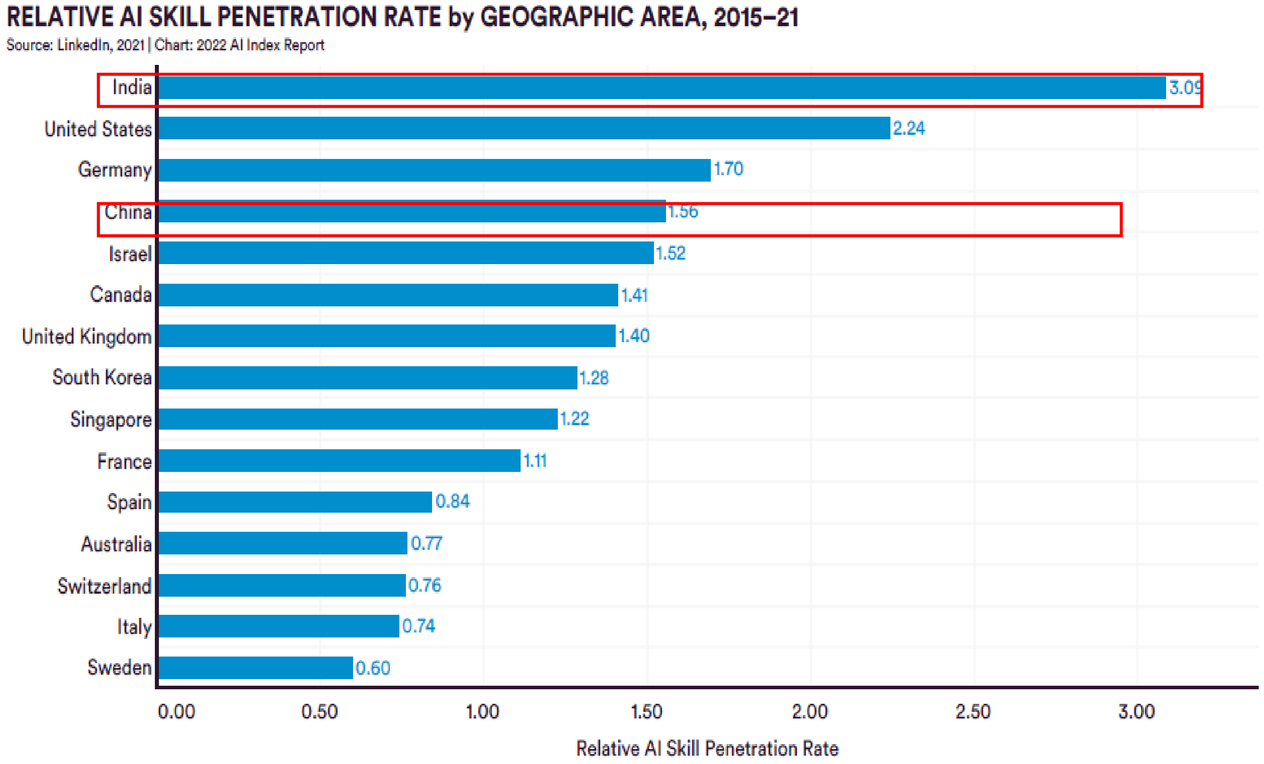
Overview

India and China are among the countries that have invested the most in AI in recent years.

The AI skill penetration rates of these two countries are also top in the world

AI skill penetration rate = *Average by Country / Average in Global*



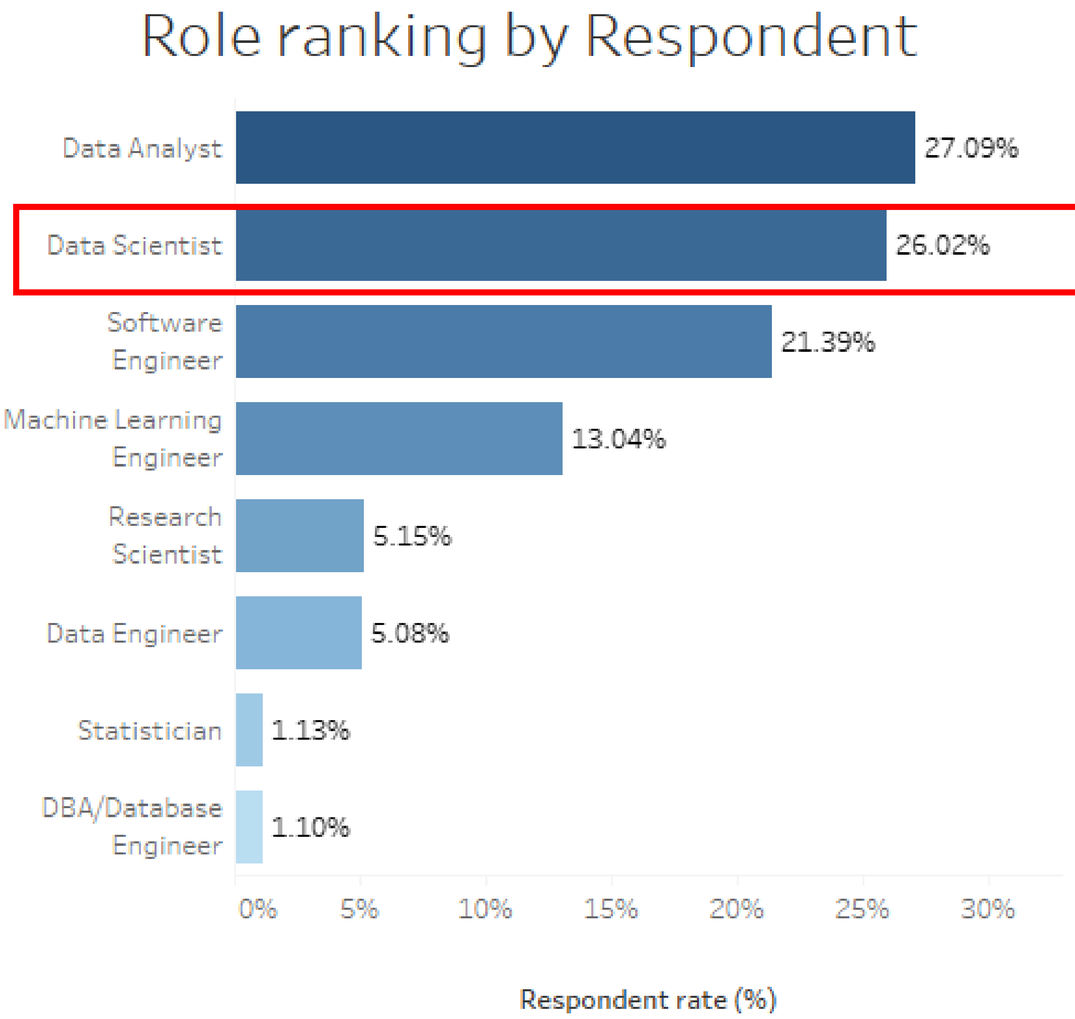


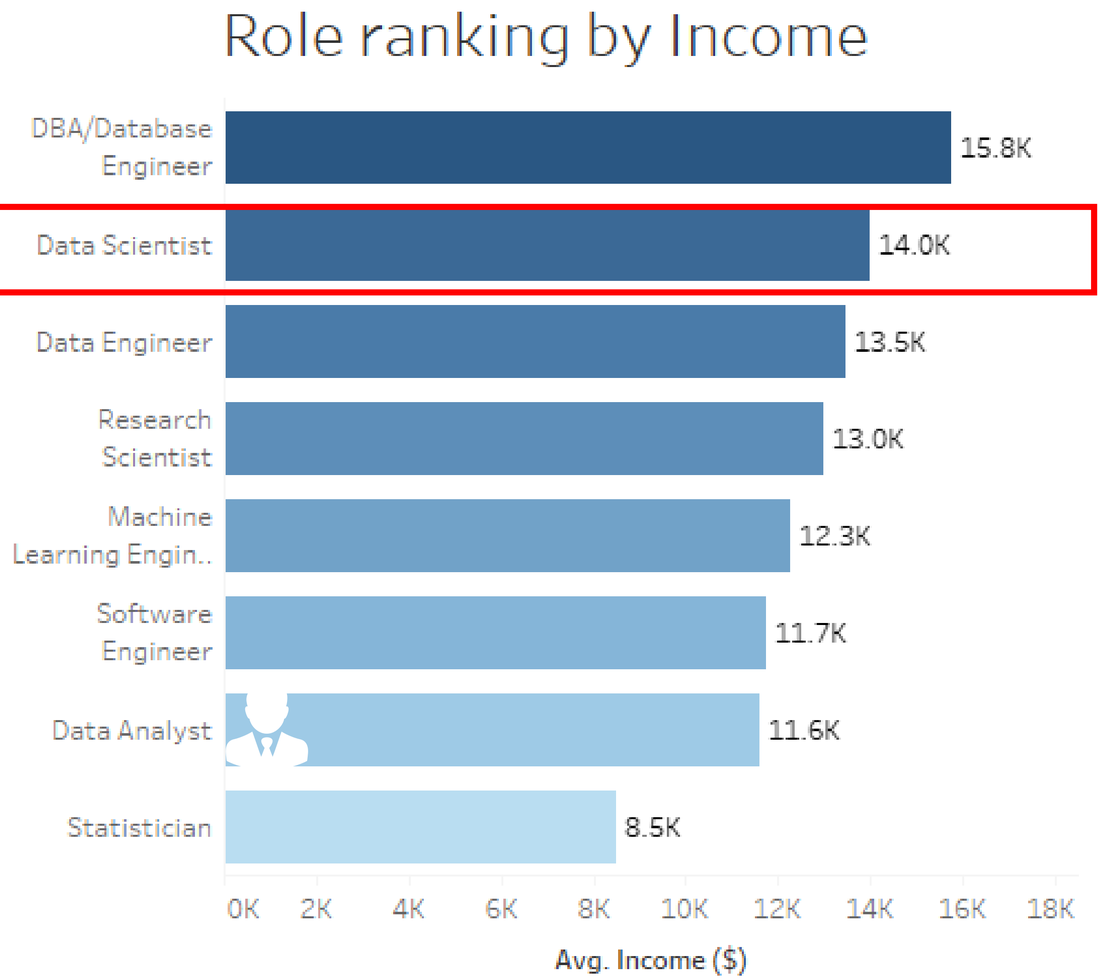
Source : Artificial Intelligence Index Report 2022 (Published by Stanford University)

Overview

Data Scientist is the hottest role here shown through:

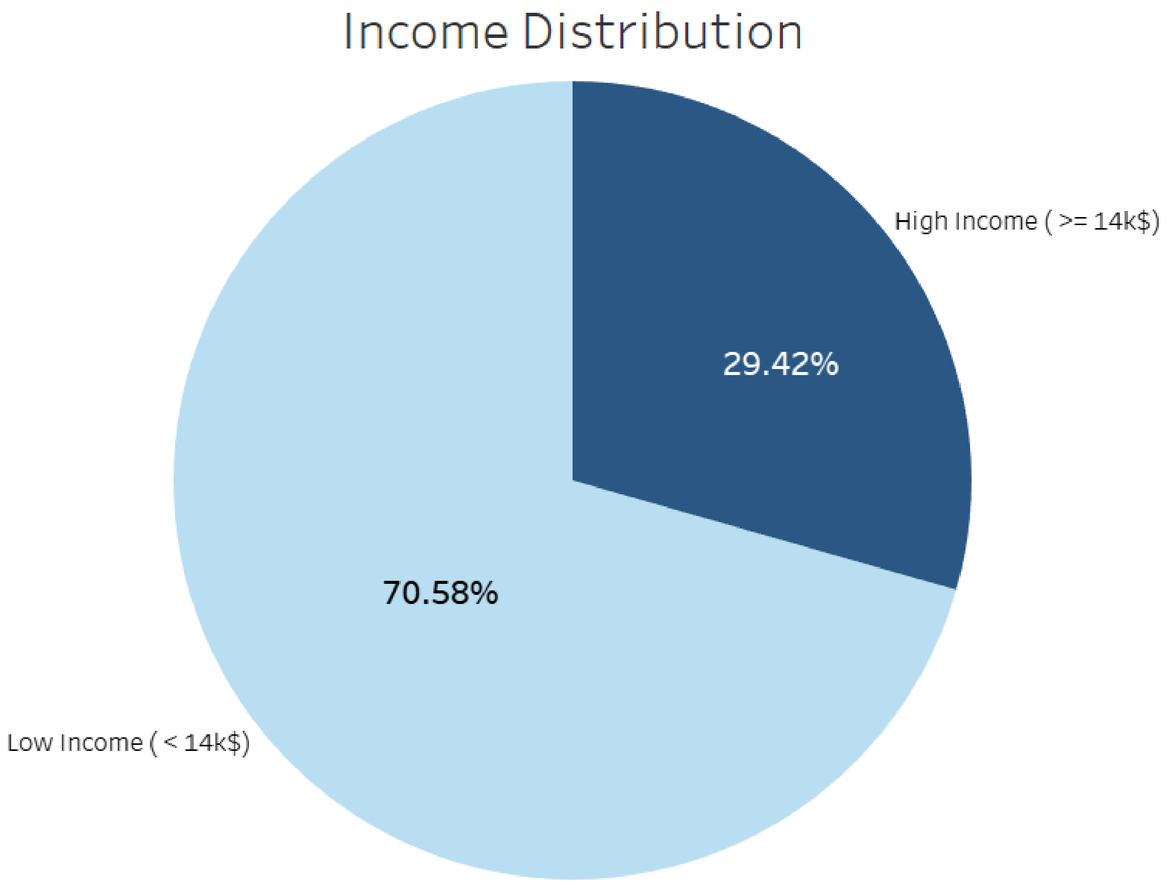
- ❖ Top 2 for Respondent rate.
- ❖ Top 2 for Income



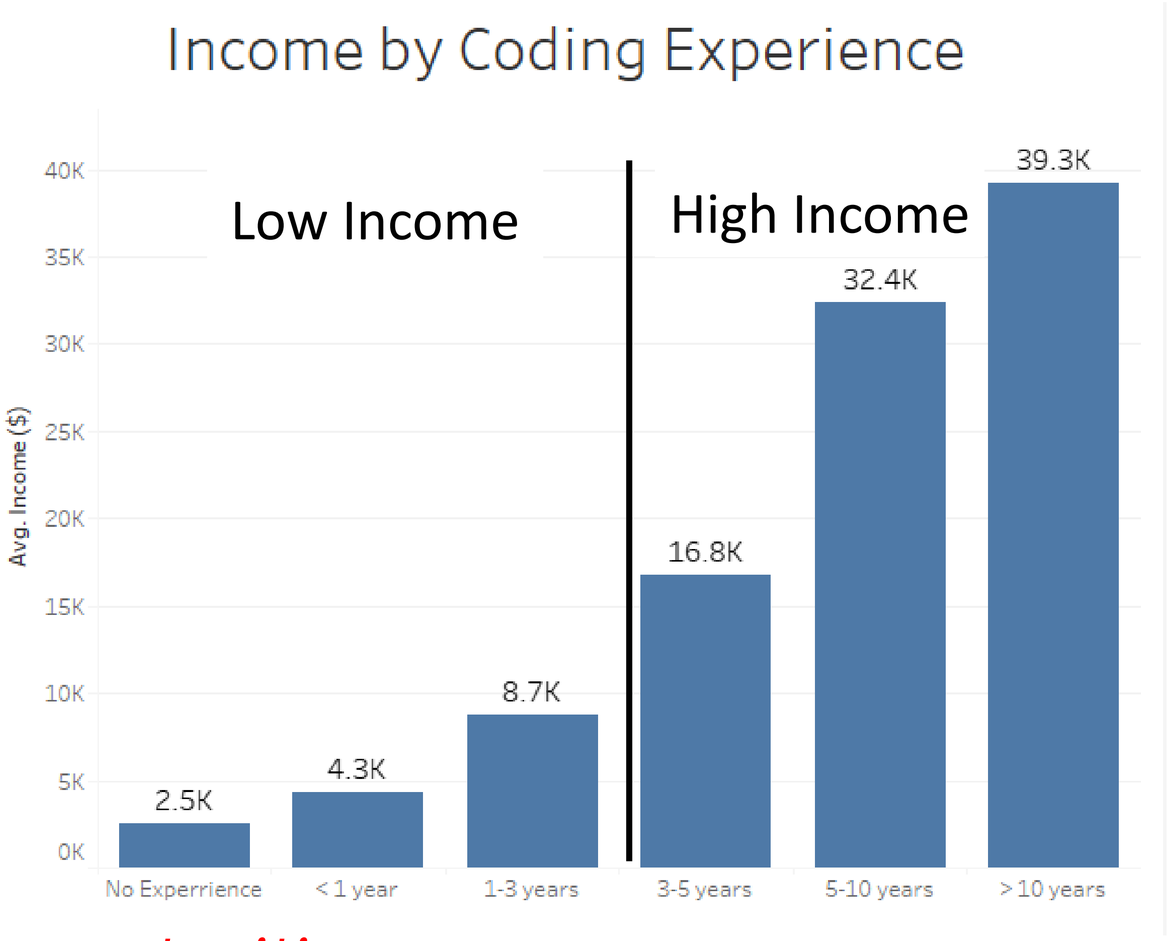


Overview

More than 70% of Respondents have incomes below the average (14k$).



This is the income equivalent to Coding Experience < 3 years.



=> This is a new role and has great growth opportunities.

Big Question

How can inexperienced Data Scientists increase their income?

Target audience
Inexperienced Data Scientists

Metric
Income

Flows of Analyst

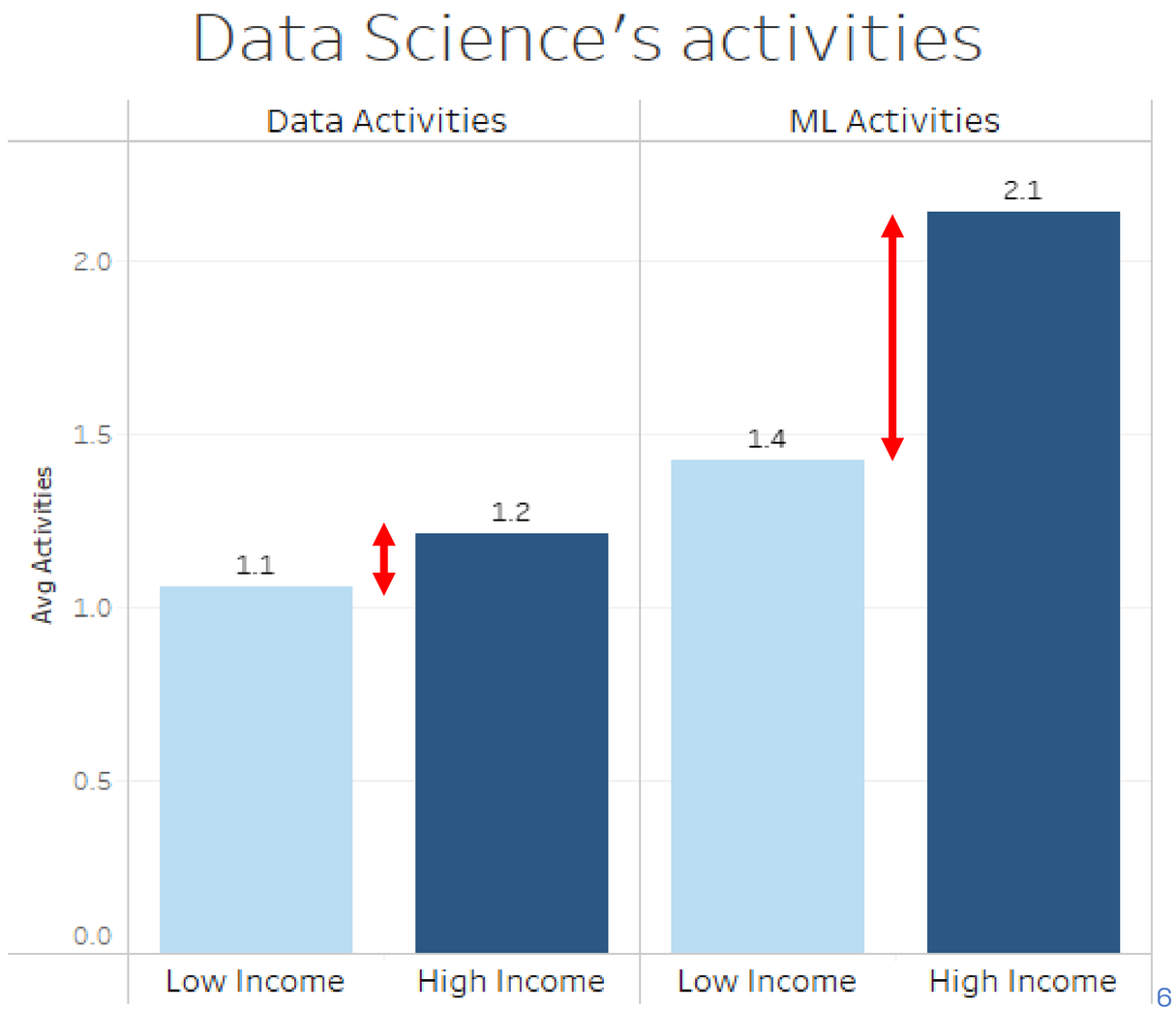
What skills impact income?	Priority skills to learn	Where to learn them?	Predict & Suggestion

What skills impact income?

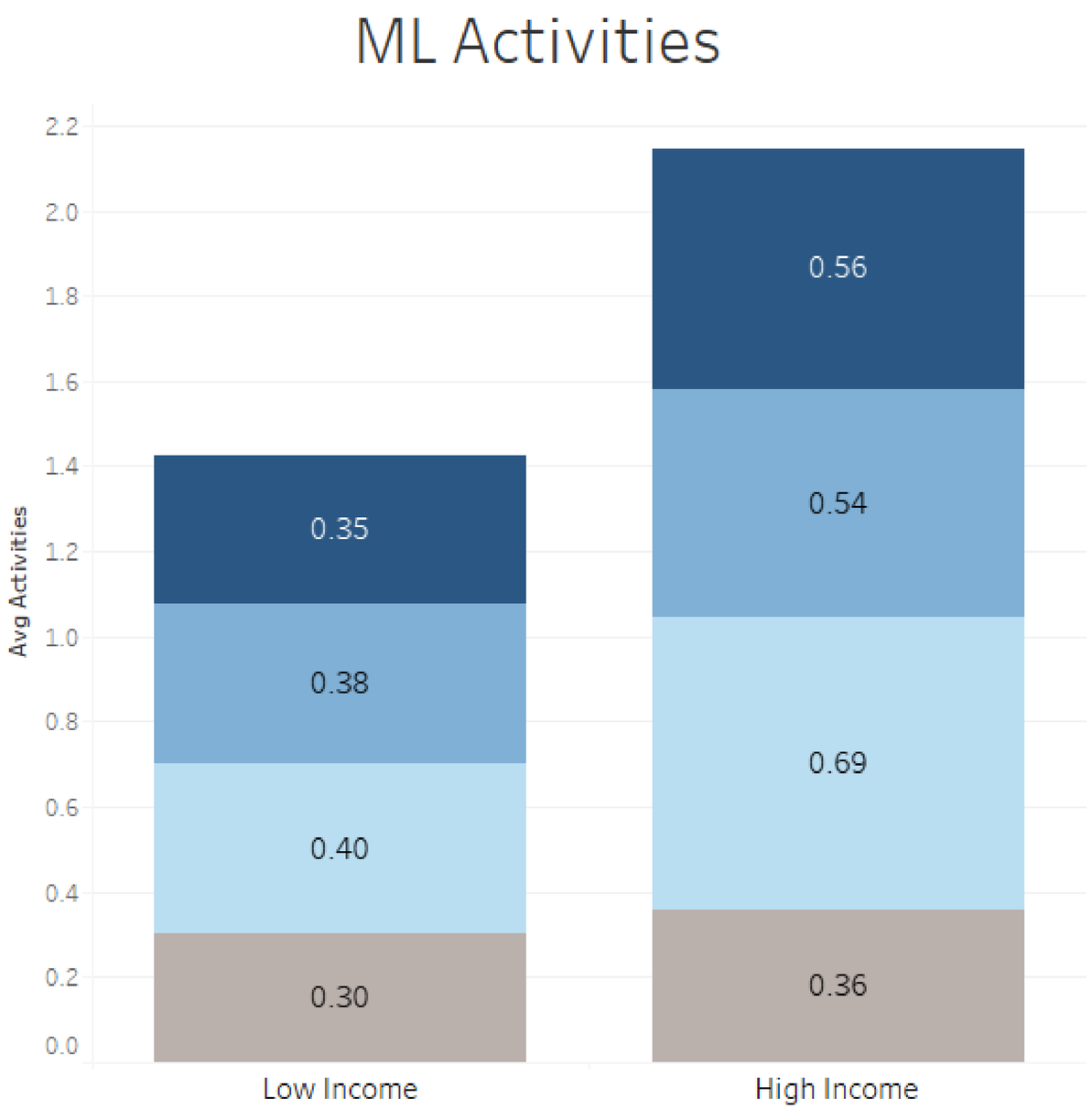
Data Scientist has 2 main groups of activities: Data and Machine Learning.

No	Activities	Type
1	Analyze and understand data to influence product or business decisions	Data Activities
2	Build and/or run the data infrastructure for storing, analyzing, and operationalizing data	Data Activities
3	Build prototypes to explore applying machine learning to new areas	ML Activities
4	Build and/or run a machine learning service that operationally improves my product or workflows	ML Activities
5	Experimentation and iteration to improve existing ML models	ML Activities
6	Do research that advances the state of the art of machine learning	ML Activities

Active participation in ML activities contributes to increasing income.



Priority skills to learn



ML activities that make the difference are:

1. Do research that advances the state of the art of machine learning

2. Build prototypes to explore applying machine learning to new areas

3. Experimentation and iteration to improve existing ML models

4. Build and/or run a machine learning service that operationally improves my product or workflows.

ML activities

- Build and/or run a machine learning service that operationally improves my product or workflows
- Experimentation and iteration to improve existing ML models
- Build prototypes to explore applying machine learning to new areas
- Do research that advances the state of the art of machine learning

Priority skills to learn

Skills focus in this report

- ML framework
- ML Algorithms

- Manage ML platforms
- ML Monitoring Tools

- Cloud computing platforms
- Cloud computing services

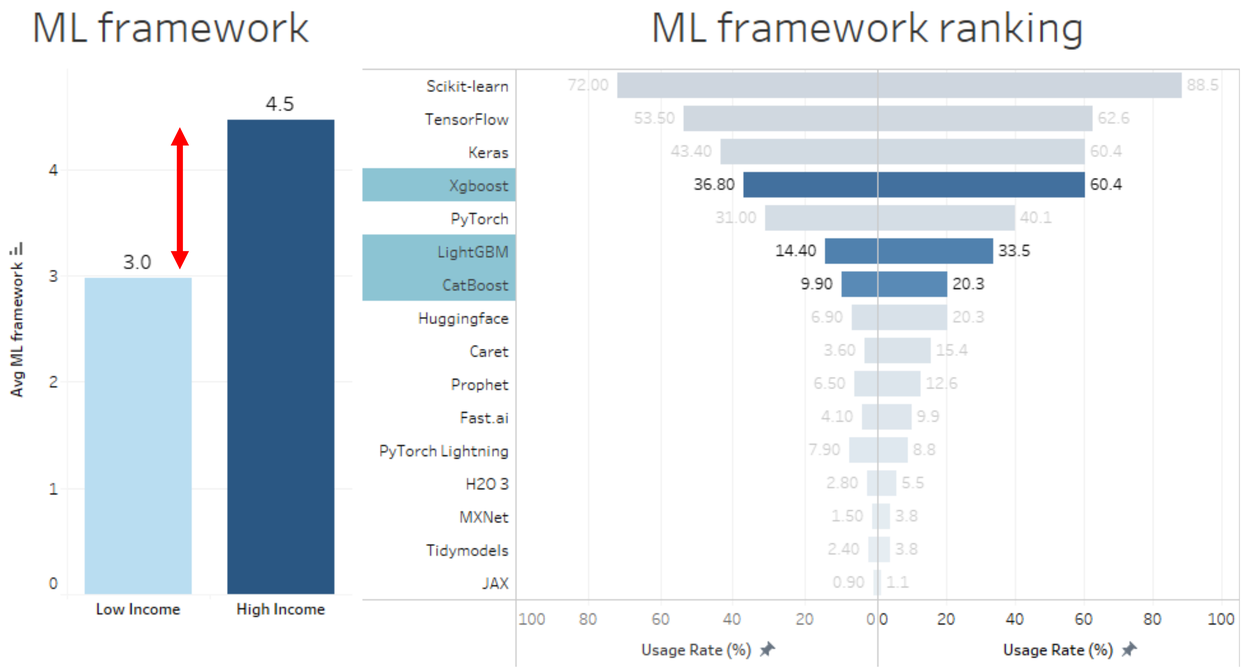
Build prototypes to explore machine learning

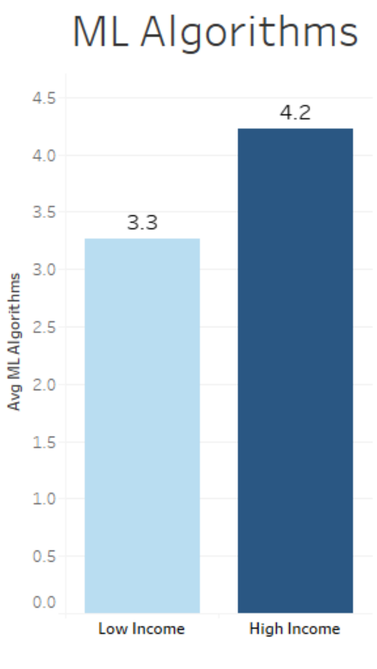
ML framework

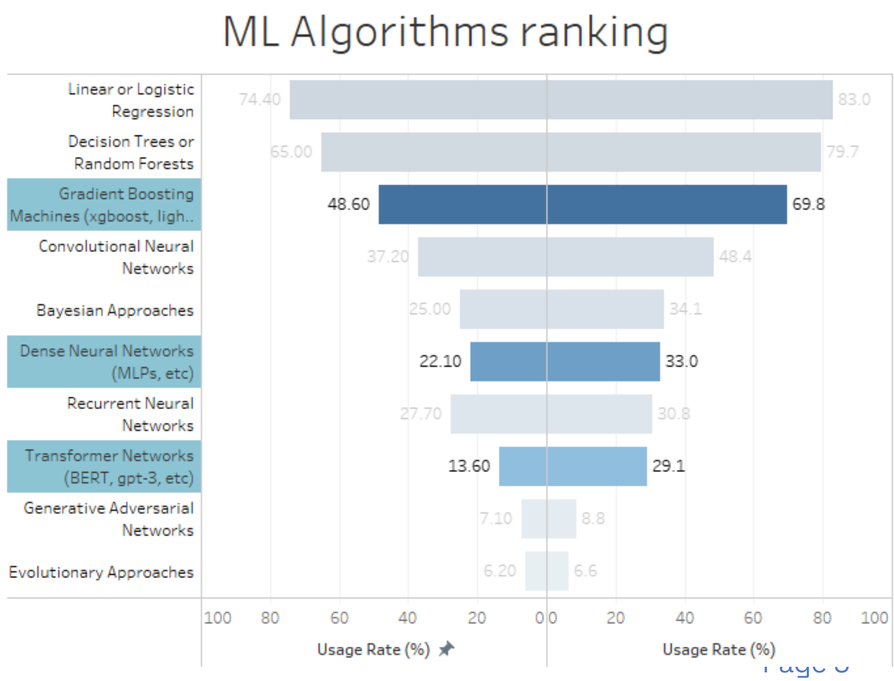
- High Income is 1.5 times more than Low Income
- This difference comes from the Xgboost, LightGBM, and CatBoost frameworks.
- These are powerful frameworks for building gradient-boosting models

ML algorithms

- Gradient-boosting algorithms have a difference, which is consistent with the left analysis
- In addition, NLP techniques (*Transformer Networks and Dense Neural Networks*) also exhibit significant differences between the two groups.







Experimentation and iteration to improve ML models

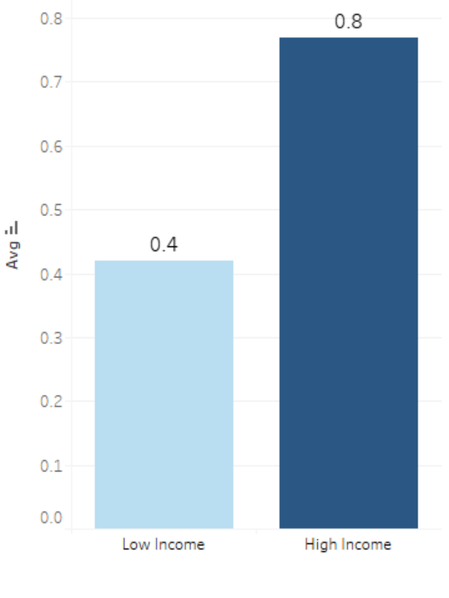
Managing ML models Top 3 platforms:

- Amazon SageMaker
- Databricks
- Azure Machine Learning Studio

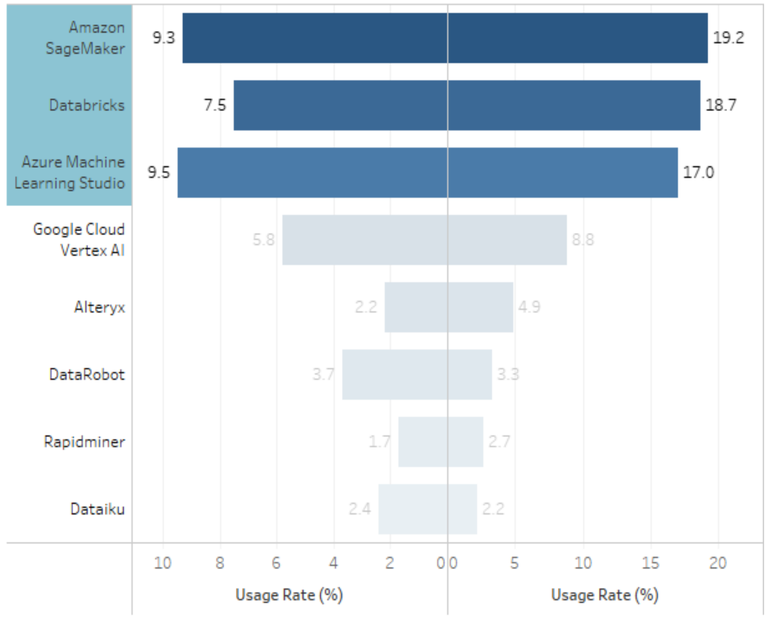
Monitoring ML models Top 3 tools :

- TensorBoard.
- MLflow.
- Weights& Biases

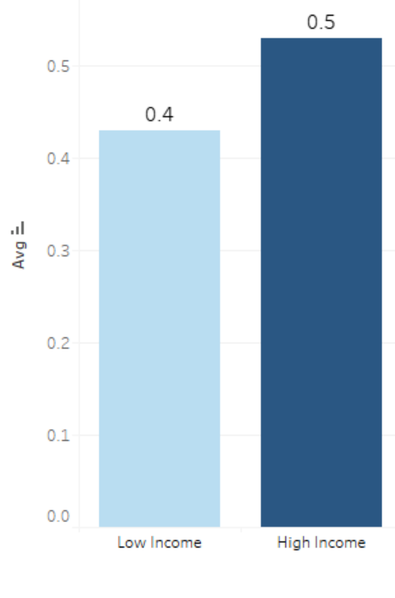
ML platforms



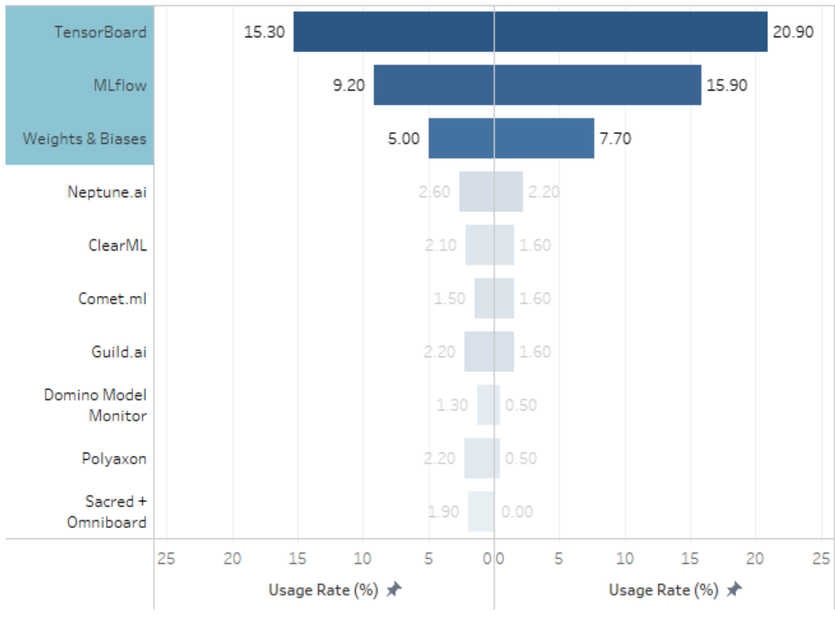
ML platforms ranking



ML monitoring Tools



ML Monitoring Tools ranking



Build and/or run a machine learning service

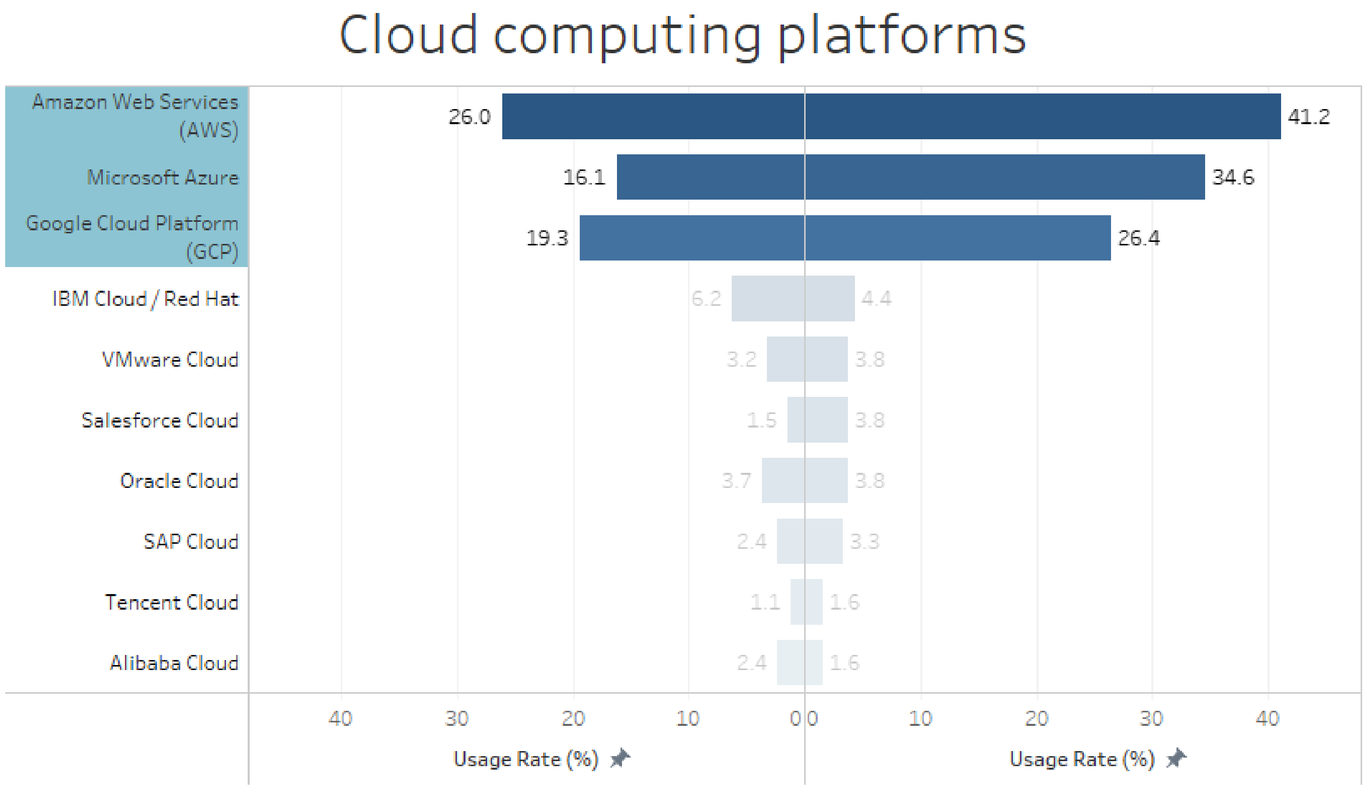
Cloud computing Top 3 platforms:

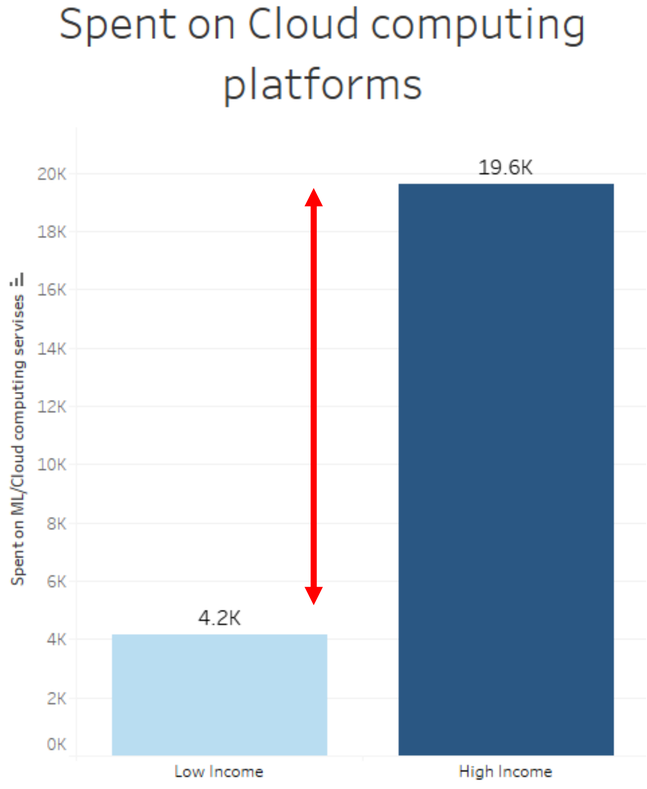
- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)

Cloud computing services

- High Income group spends almost 5 times more than the Low Income group.

They have experience in selecting and optimizing cloud computing services => increase efficiency and reduce costs for the company.





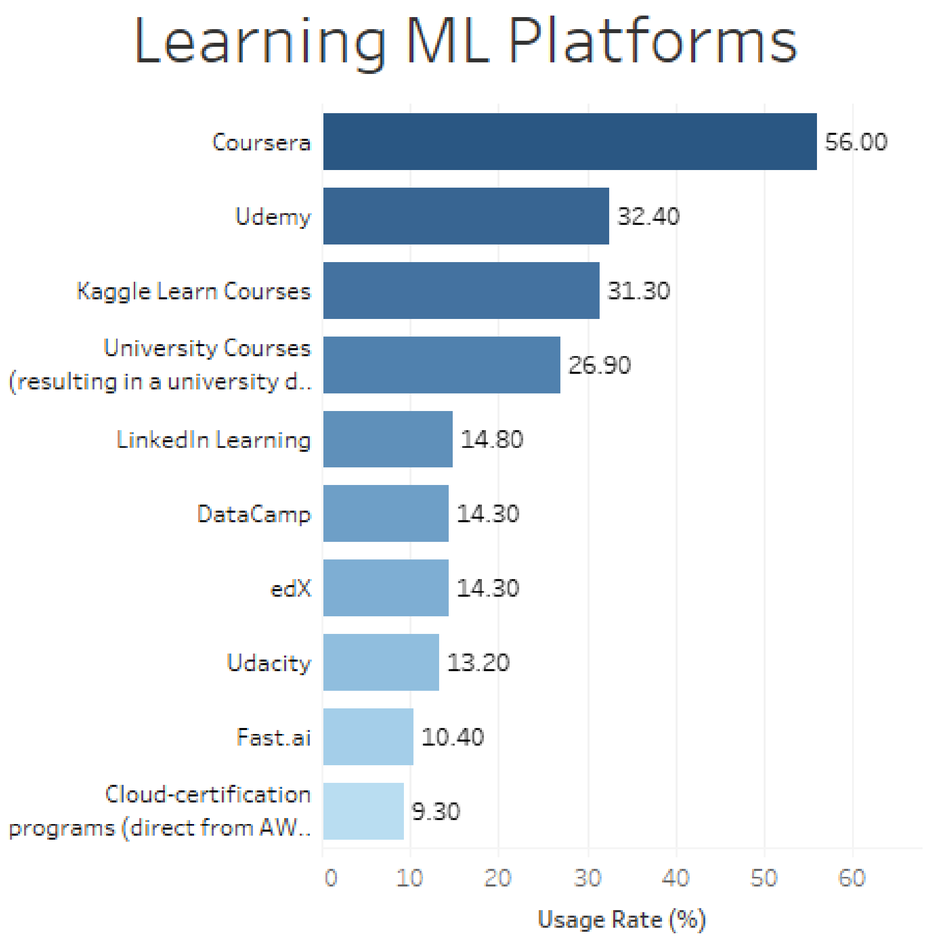
Where can I learn ML ?

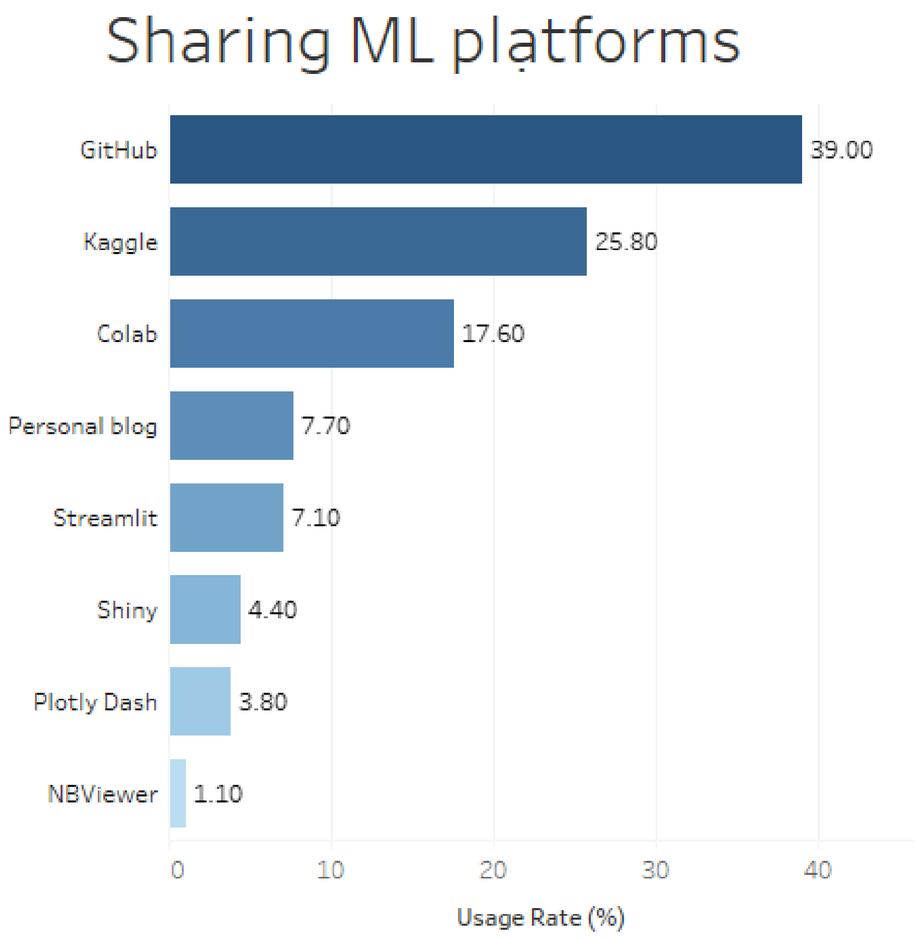
Top 3 **Learning platform** :

- Coursera
- Udemy
- Kaggle Learn Courses

Top 3 **Sharing ML platform** :

- GitHub.
- Kaggle.
- Colab.





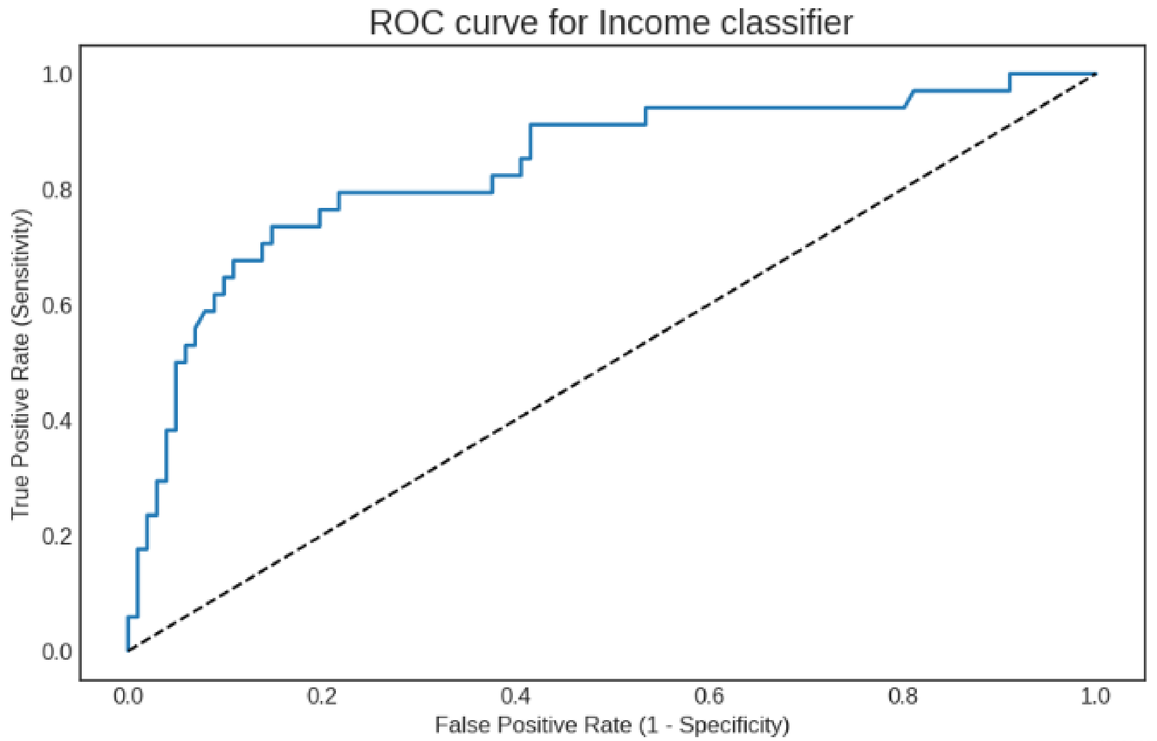
Model Predict Income

The logistic regression model accuracy score is 0.8370. So, the model does a very good job of predicting Income level of Data Scientist

ROC AUC of our model approaches towards 1. So, we can conclude that our classifier does a good job of predicting Income level of Data Scientist

The accuracy of the original model test and GridSearch CV are both 0.8370 => No improvement.

	precision	recall	f1-score	support
0	0.88	0.91	0.89	101
1	0.70	0.62	0.66	34
accuracy			0.84	135
macro avg	0.79	0.76	0.77	135
weighted avg	0.83	0.84	0.83	135



Recommend skill need to prioritize improvement

```python
def find_similar_index(df, target_index):
    target_row = df.iloc[target_index]

    filtered_df = df[df['Income level'] == 2]

    similarities = []
    for _, row in filtered_df.iterrows():
        similarity = np.linalg.norm(row.values[2:] - target_row.values[2:])
        similarities.append(similarity)

    most_similar_index = similarities.index(min(similarities))
    most_similar_row = filtered_df.iloc[most_similar_index]

    improvements = np.where(most_similar_row.values > target_row.values, most_similar_row.values, target_row.values)

    columns = df.columns[2:]

    improvements_df = pd.DataFrame([target_row.values[2:], improvements[2:]], index=['Initial', 'Recommend'], columns=columns)

    return improvements_df
```

Example

```python
improvements = find_similar_index(pre_df,9)
improvements
```

	Coding Experience	Spent on ML	Used TPU	ML framework	ML algo	ML platforms	ML monitor
Initial	0.5	500.0	3.5	0.0	1.0	2.0	0.0
Recommend	2.0	500.0	3.5	3.0	2.0	2.0	0.0

Conclusion

- To increase income, inexperienced Data Scientists need to improve their skills to do more Machine Learning activities.
- Priority skills :
  - ✓ ML framework: Xgboost, LightGBM, and CatBoost.
  - ✓ ML algorithms Gradient-boosting, NLP techniques.
  - ✓ Managing ML models: Amazon SageMaker, Databricks, Azure Machine Learning Studio
  - ✓ Monitoring ML models: TensorBoard, MLflow, Weights& Biases.
  - ✓ Cloud computing: Amazon Web Services, Microsoft Azure, Google Cloud Platform.
- Learning & Sharing platforms: CourseraUdemy, Kaggle, Learn Courses, GitHub, Colab,...

Thanks for Watching