NYU | Center for Data Science

# StrongHires

# Capstone Project - DS-GA 1001

Sargun Nagpal (sn3250)

Harsha Koneru (hk3820)

Sharad Dargan (sd5251)

Author Contributions:
Sargun Nagpal - Data Prep, Hypothesis Testing and Regression
Harsha Koneru - Data Prep, Hypothesis Testing and Clustering
Sharad Dargan - Data Prep, Hypothesis Testing and Classification

N-Number used as seed: **12132707 (Sargun Nagpal)**

**Code Repository Link:** Github

## Table of Contents

# 1. Introduction

The world of data science is highly diverse and rapidly evolving. For data science practitioners, this means they have to stay abreast of new developments and methodologies in order to stay relevant. The evolution of the field is creating new specializations which didn't exist before, and old jobs might at some point in the near future become obsolete. On one hand, there is a new benchmark being broken, or a new boundary being pushed every other week, and on the other there is frequent discussion about a "Data Science Winter", which involves a significant reduction in trust from the ecosystem on the efficacy and usefulness of data science.

In this project, we aim to cut through some of the noise, to answer two broad questions:
1. What skills and education are important for what specializations in Data Science?
2. How do skills and industries impact how much we earn as Data Science practitioners?

The onset of the economic slowdown and the recent job losses have made these questions even more pertinent. While there is a lot of anecdotal information and great advice available online for people who are just entering the field, we felt there is a need for using Data Science to help Data Scientists.

## 1.1 The Dataset

The Kaggle Data Science Survey is a survey conducted by Data Science Competition platform Kaggle to understand the state of Data Science and Machine Learning every year. This survey has questions pertaining to the education levels, the work experience, and the tools data science practitioners use to solve problems.

For this project, we worked with survey responses from the years 2020[1], 2021[2] and 2022[3]. The respondents to these surveys are users of the Kaggle platform, and are data science practitioners. To give context about the demographics of the survey respondents, the 2022 survey contained almost 24,000 respondents, with 77% being male and 22% being female. 36% of the respondents are from India, around 12% from the USA and Brazil comes in third with ~3.5% of the respondents. Roughly 55% of the respondents are below the age of 30. 38% of the respondents have a master's degree, 31% have a bachelor's degree and 11% have a PhD. Around 42% of the respondents have claimed to have published academic papers. The three most common programming languages used are Python, SQL and R.

This dataset across the 3 years has around 578 columns representing 44 questions (Single choice and MCQ) pertaining to work experience, industry, education, tools and frameworks used, cloud computing services used, hardware used, and algorithms used and 70,000 rows, each representing individual survey reponses from the participants

**Note:** The questions have changed slightly across the years and we have removed questions which were not present in all three years.

## 1.2 Data Pre-Processing

The Data Pre-Processing was split into four main parts:

1. **Combining survey data of 3 years**
   a. Data was gathered from three years of surveys. Thus, there is a temporal structure to the data. There are also slight differences in the number of questions asked in each version of the survey.
   b. We mapped the questions across 3 years and standardized them.
   c. Added a unique respondent identifier for later consumption in feature engineering.
   d. Merged the datasets to create a master dataset for consumption across regression, clustering and classification tasks.
2. **Handling Missing Data:** Missing responses represent that the participant does not identify as having the skill/working on the technology. Hence, Missing responses for questions were replaced with 0 for Binary and MCQ Questions, and with "Unknown" for Categorical variables.
3. **Handling Inconsistent Values:** Removed survey responses in which the person is not more than 29 years old and claims the number of years of Programming/ML experience as over 20
4. **Handling Survey responses and standardizing them over the three years**
   a. For each Multiple Option Question, we created multiple columns containing the responses of the respondent. For example, if the question was about the Machine Learning libraries used by the respondent with the options being Tensorflow, SKLearn, Keras etc, we created columns for each of the libraries and had 1 if the user marked that they used the library and 0 if they marked that they didn't.

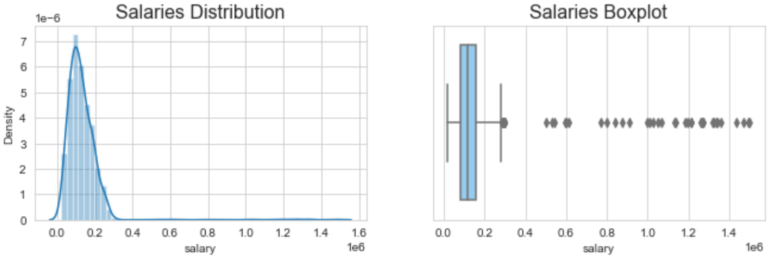
b. In some cases, the options were different. For example, in a question asking for programming experience, in one year, the option was “1-2 years” and in another year, the option was “1-3 years”. We dealt with this on a case by case basis, merging the options when we deemed it necessary.

5. **Gathering and incorporating salary data for the prediction (regression) task**

a. To convert the salaries to a numerical scale, we downloaded a 3rd party dataset of Data Science salaries in the United States and used Inverse Transform Sampling to sample a numerical salary for each salary range based on the distribution of salaries within that range in the third party dataset.

b. Note that the benefit of this approach (over uniformly sampling from a salary interval) is that the distribution of salaries in an interval may not be uniformly distributed. For example, for the salary range 200-250k, it is much more likely to have a salary of 200k.

c. The figure attached shows the distribution of salaries after converting them to a numerical scale. It can be observed that some individuals have very high salaries, observed as outliers in the boxplot.



6. **Feature Engineering for the Classification, Clustering and Regression task**

a. The features we prepared can be grouped into the following categories:

i. **Job Title:** Data Scientist, Data Analyst, Data Engineer, ML Engineer, Research Scientist.

ii. **Skills and Experience:** Programming experience, Uses Python/ SQL/ R/ AWS/ Tensorflow etc, No. of programming languages / ML frameworks/ BI Tools known.

iii. **Industry Information**: Industry- Tech, Finance, Pharma etc, Data science team size.

iv. **Personal Information:** Gender, Level of education (Bachelor’s, Master’s, PhD).

b. Methodology for Encoding:

i. The **Ordinal variables** such as Age, Education, Programming/ML Experience were encoded using Ordinal Encoding to preserve the sense of ordering of level in the variables

ii. The **Binary Responses** (e.g. Used TPU?) were encoded with one-hot encoding, since captures the binary nature of the responses and yet doesn’t increase the dimensionality of the dataset. We end up with 36 such features

iii. The **MCQ Responses** were handled in 2 major ways:

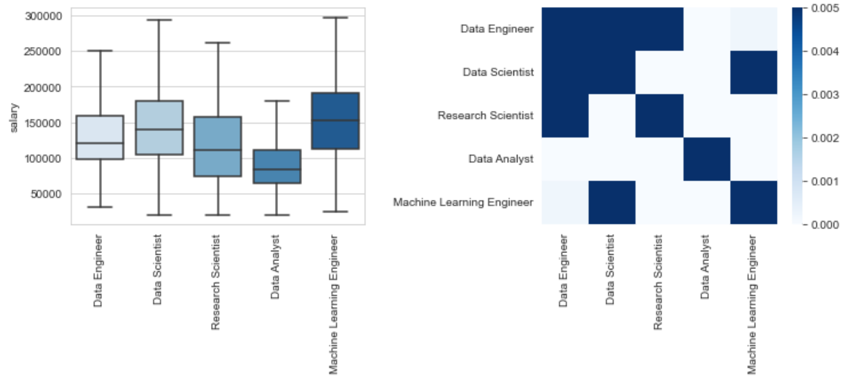
1. The questions with a high number of levels were aggregated to reflect the degree of experience/engagement with the question subject matter. e.g. Responses to “Which Visualization libraries are you familiar with” were aggregated to reflect the “Num of Viz Libraries user is familiar with”. We have 15 numerical features from this
2. Secondly, some levels of in the MCQ responses were extracted as individual features, e.g. - “Used Python” was extracted from “What programming languages are you familiar with”

The entire dataset, and the feature engineered dataset have both been added to the project file, and pushed to our github repository.

## 2. Hypothesis Testing: “Studying at CDS is expensive! Will I get my returns?”

### 2.1 What jobs will earn you the most money?

- **Methodology & Reasoning**: We filtered the salaries of five job titles in the US (Data Engineers, Data Scientists, Research Scientists, Data Analysts and ML Engineers) and performed the non-parametric Kruskal-Wallis H test to check if the medians of the groups have a significant difference. Thereafter, we performed Dunn's post hoc test with Bonferroni correction (to account for multiple comparisons) to identify pairs of titles that are significantly different. We plotted a heatmap with an upper limit of alpha (0.005), so as to highlight significant results with a light blue shade. We have high



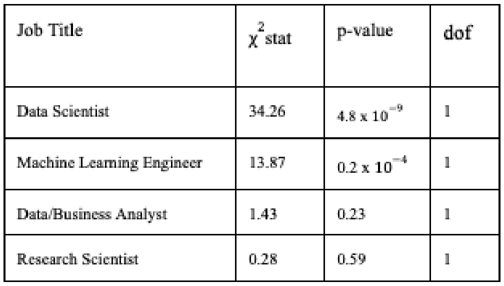
sample size (n~2,400), hence the study should have a reasonably high power

- **Findings & Conclusion**: Salaries differ significantly by job titles ( *Kruskal Wallis* $H = 468$, $dof = 4$, $p\text{-value} = 5.58\text{x}10^{-100}$). ML Engineers have the highest median salary (>150k), significantly different from all other titles except Data Scientists (Heatmap: $p < 0.005$). Data Analysts have the least median salary (<100k), significantly different from all other job titles. Therefore, this analysis suggests that if it is money that you are looking for, being an ML Engineer or a Data Scientist could be your best bet.

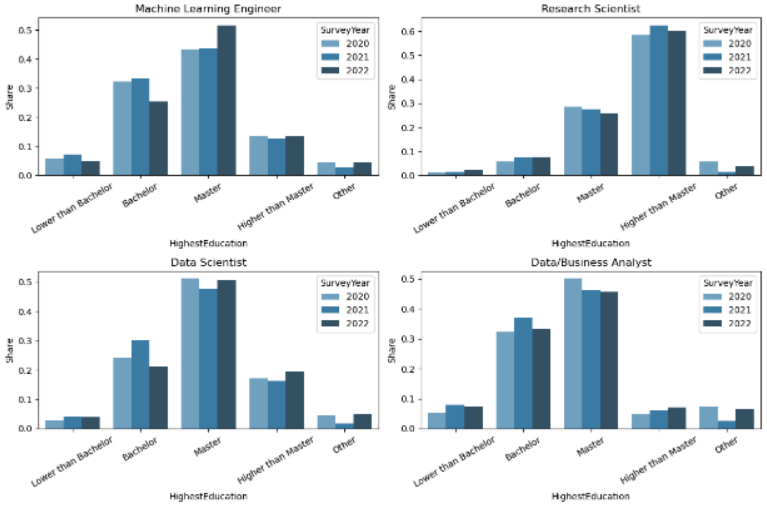
### 2.2 Is the representation of Advanced degrees (Masters & above) among Data Professionals increasing over years?

- **Methodology & Reasoning**: We take a subset of the survey data of three years (2020-2022) for 4 job titles (Data Scientist, Data Analyst, Machine Learning Engineer, Research Scientist) and performed the $\chi^2$test to check if the highest education level is different across the 3 years. Since, we are testing categorical variables for independence across 3 groups, we use the $\chi^2$test. We have high sample size (n~22,000), hence the study should have a reasonably high power

- **Findings & Conclusion**: The education level for the different roles across the 3 years is significantly different (alpha:0.005) for Data Scientist and Machine Learning Engineer with the highest shift in the share of Master's degree for Machine Learning Engineers. Hence, even though you might get more money being an ML Engineer or Data Scientist, you need to invest in a master's first. CDS might not be a bad idea after all!



| Job Title | $\chi^2$ stat | p-value | dof |
|---|---|---|---|
| Data Scientist | 34.26 | $4.8 \times 10^{-9}$ | 1 |
| Machine Learning Engineer | 13.87 | $0.2 \times 10^{-4}$ | 1 |
| Data/Business Analyst | 1.43 | 0.23 | 1 |
| Research Scientist | 0.28 | 0.59 | 1 |



## 3. Regression: "Here is my data science profile. How much should I earn?"

**Question:** Can we predict data science job salaries of individuals in the United States based on covariates such as their job title, industry, skill sets and experience ?

### 3.1 Data Preparation

Data Preparation and Feature Engineering was done (as explained in Section 1.2)

The target variable (salary) was on a categorical scale, and was converted to a numerical scale by using Inverse Transform Sampling on a 3rd party dataset [4] of Data Science salaries in the United States to sample a numerical salary for each salary range.

### 3.2 Modeling

#### 3.2.1 Train-test split

We did a 80:20 random split of the dataset (Train ~ 1900 recs, Test ~ 500 recs) and compared the distribution of the train and test set salaries to ensure that we have a representative split. We also created an adjusted dataset by removing outliers with salaries > 500k. We trained and evaluated our models on both the original and the adjusted dataset separately.

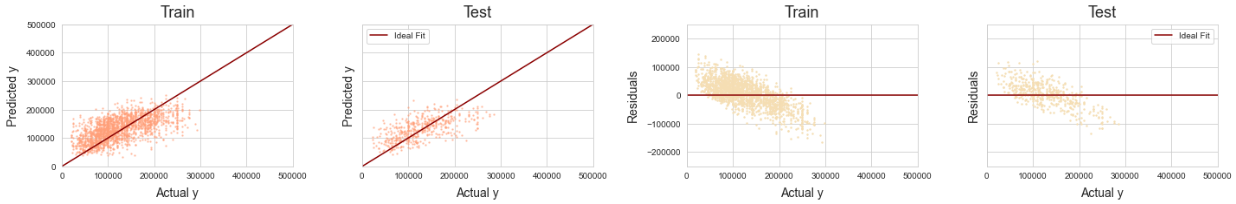
### 3.2.2 Performance Metrics

We aimed to minimize the RMSE of our model, but found out that it is hard to predict the exact salary values using our features, so we shifted to maximizing the $R^2$, which is the square of the Pearson correlation coefficient between our predictions and the true values. A model with a high $R^2$ score is useful because if it predicts a high salary, the actual salary is expected to be high too (and vice versa). We also calculated other metrics to assess the goodness of fit: MAE, MAPE and the Maximum Error.

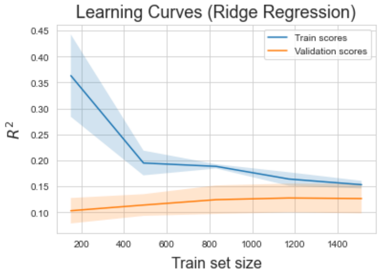
### 3.2.3 Cross-validation strategy & Hyperparameter tuning

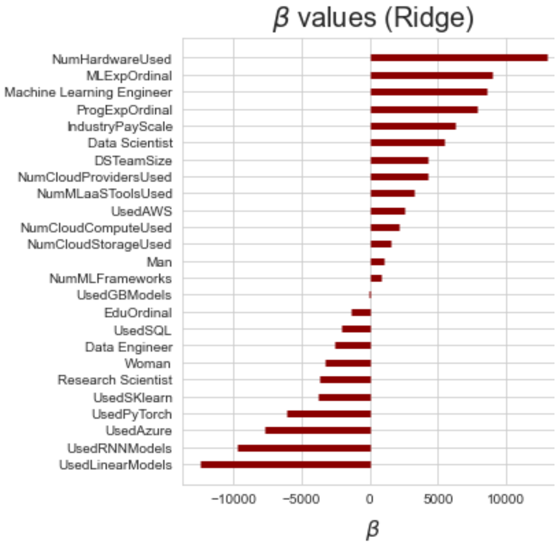
For each ML model, we defined a range of values of relevant hyperparameters and performed a 5-fold cross validation with Randomized search to search for the best set of hyperparameters. The best model was evaluated on the train and test set.

### 3.2.4. Best Model (Ridge Regression)



- The plot of the actual vs predicted values shows that the model is making meaningful predictions, but often overestimates or underestimates the true salaries.
- The residual plot reveals that we are over-estimating salaries on the lower range, and under-estimating salaries on the higher range. The residuals seem to be negatively correlated to the salaries, and are not normally distributed.
- The learning curves reveal that both the Train and Validation set $R^2$ values converge, but both are low, indicating that the model is underfitting.
- The beta values (coefficients) of the model reveal that features such as being a ML Engineer or having higher programming experience (while controlling for the other variables) relate to higher salaries, while using Linear models, RNNs, or Azure is related to lower salaries.
- Unfortunately, being a man has a positive coefficient, while being a woman has a negative one. However, since this is an under-fitted model, we should be careful about interpreting beta values.





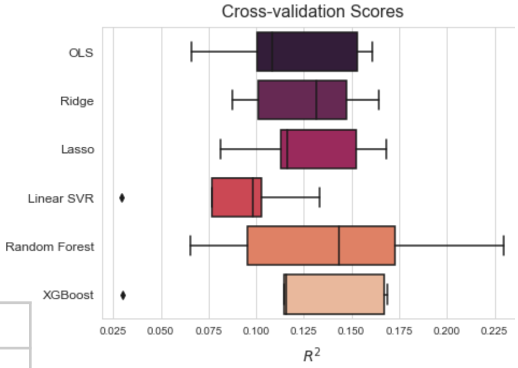
| | Train | Test | Adj Train | Adj Test |
|---|---|---|---|---|
| **R²** | 0.151 | 0.186 | 0.388 | 0.393 |
| **RMSE** | 102521.84 | 101317.698 | 42944.219 | 42440.21 |
| **MAE** | 44011.925 | 41876.663 | 33310.749 | 32443.283 |
| **MAPE** | 0.395 | 0.358 | 0.346 | 0.323 |
| **Max Error** | 1367851.779 | 1214754.456 | 164553.87 | 141881.174 |

- The $R^2$ value on the adjusted test set is about 39% (much higher than just predicting the mean, which would yield an $R^2$ value of 0).
- The RMSE value on the adjusted test set is about 42k, which means that the model is just useful as a salary tier predictor, and not the exact salary value.

### 3.2.5 Model Comparison

Random Forest has the highest median cross-validation score, but very high variance, indicating overfitting. All models have low $R^2$ scores, indicating that the predictive power of the features for salary prediction is low.

Ridge generalizes the best on the full test set, while XGBoost performs the best on the Adjusted test set, yielding an $R^2$ of 0.40. The systematic patterns observed in the residual plots indicate that a generalized linear model may be more suitable for this problem. Indeed, a Poisson GLM model fitted on the dataset performs the best on the overall test set. Further work is required to explore the performance of Generalized linear models with suitable hyperparameter tuning.



| | Model | Train $R^2$ | Test $R^2$ | Adj Train $R^2$ | Adj Test $R^2$ |
|---|---|---|---|---|---|
| 1 | OLS | 0.159 | 0.18 | 0.389 | 0.393 |
| 2 | Ridge | 0.151 | **0.186** | 0.388 | 0.393 |
| 3 | Lasso | 0.157 | 0.184 | 0.386 | 0.392 |
| 4 | Linear SVR | 0.092 | 0.113 | 0.306 | 0.298 |
| 5 | Random Forest | 0.863 | 0.165 | 0.674 | 0.399 |
| 6 | XGBoost | 0.433 | 0.178 | 0.516 | **0.404** |
| 7 | Poisson GLM | 0.182 | **0.223** | 0.396 | **0.401** |

## 4. Clustering: "Can we cluster respondents into different job roles based on their skills?"

Since the goal of the classification task is to try and infer the most relevant kind of data practitioner (Data Analyst, Data Scientist, Machine Learning Engineer etc) jobs a candidate qualifies for from the skill sets reported, we decided to cluster these practitioners based on their skill sets, and to check if the practitioners are separable based on them. We tried two different approaches:

1. Traditional KMeans clustering using only numerical data.
2. KPrototypes clustering, which is compatible with categorical information as described by Huang (1998)[5].

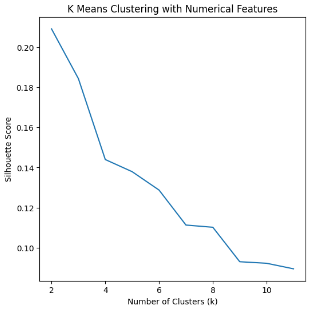
One significant challenge in performing unsupervised clustering using our data, is that most of the information in our dataset is categorical, and can't be interpreted in the form of means. This means traditional approaches like K-Means Clustering would have limited success probability.

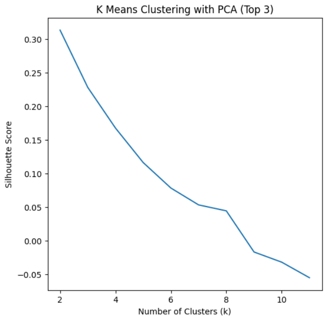
### 4.2 Data Preparation

Data Preparation and Feature Engineering was done (as explained in Section 1.2)

### 4.3 KMeans Clustering

Since most of our predictive features are categorical, K Means yielded less than ideal results. It was trained solely on the numerical features of the dataset. The best silhouette score using this approach was 0.23. Even when computed on a dimensionally reduced dataset after Principal Component Analysis, the best silhouette score was only 0.32. Figures attached show the silhouette score plots of both approaches:



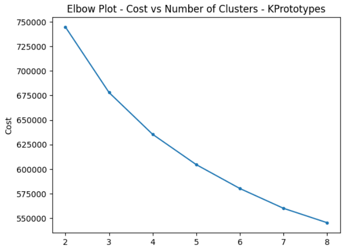


### 4.3 KPrototypes Clustering

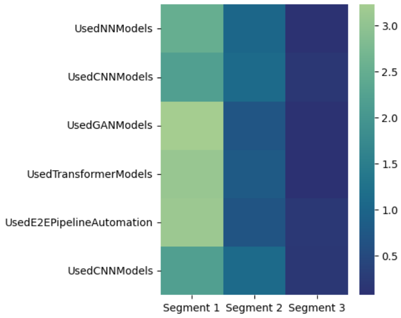
The K-Prototypes Algorithm is a clustering algorithm which integrates the  K-means and K-modes clustering algorithm and enables us to work with mixed data types. It does this by defining a combined dissimilarity measure which integrates information from both categorical and numerical features. Using this approach, we found interesting and interpretable results from our clusters.

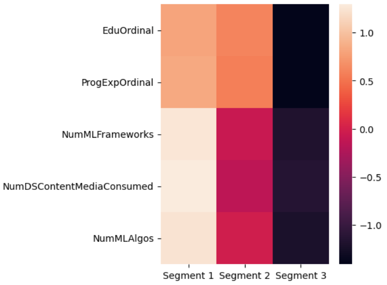
#### 4.3.1 Summary of experiments and results

Since a silhouette score is based on distances, it is not directly applicable to data with categorical features. Thus, the elbow method was used. Using the elbow plot (atttached), we identified that the ideal number of clusters could be either 3 or 4. Upon further investigation and analysis of the clusters, we settled on 3 clusters as the ideal number of clusters.



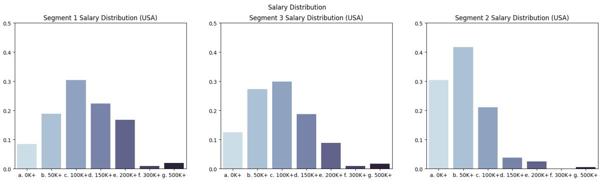
**We found that the three clusters represented the modernness and sophistication of data practitioners and not the various kinds of data practitioners (as we had previously expected).** The heatmap plots attached show a selection of features and their normalized proportions/means for a handful of categorical and numerical features respectively.



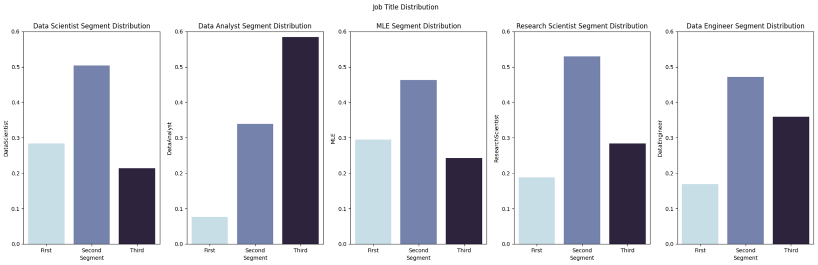


The key observations from the cluster analysis are listed below:

- The people in Segment 1 seem to be working with the cutting edge in machine learning and data science. They have a higher proportion of people who have been using modern techniques like GANs, Transformers, Convolutional Neural Networks etc. They also have significantly higher end-to-end ML pipeline automation experience.
- From segment 1 to segment 3, we see a reduction in the number of frameworks and algorithms the practitioners work with. However, we see no difference in the highest level of formal education and programming experience between segments 1 and 2.
- In line with the kind of work and the skill level, Segment 1 seems to be commanding more salaries on average as compared to the other two segments. As we progress from segments 1 to 3, we see that the salaries get more and more left skewed.



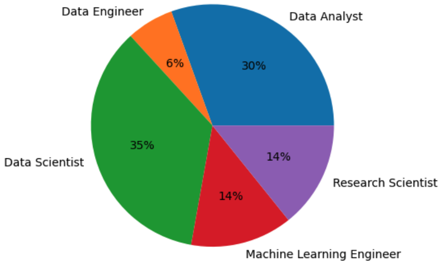
- However, except for in the case of data analysts, there is no distribution of specific professions in specific segments. This indicates that job titles are very fluid, and since the field is still nascent, it will take more time for these titles to take on concrete meanings which are separable.



## 5. Classification: Identify the most suitable jobs for a user based on the user's responses about their skills, exposure and experience.

### 5.1 Data Preparation

**5.1.1 Target Variable:** Identified 5 major job titles in the Industry that have a reasonable difference in their Job Descriptions. Data is filtered to include only these five classes in the target variable (Job Title). The target variable is label encoded to encode the 5 classes - Data Scientist, Data Analyst, Machine Learning Engineer,



Business Analyst and Research Scientist. There is a class imbalance in the target variable (figure attached).

**5.1.2 Feature Engineering:** Data Preparation and Feature Engineering was done (as explained in Section 1.2). In order to shortlist a smaller number of features that help in splitting the classes as effectively as possible, we performed feature selection using mutual information score, and based on this we picked 30 most important features.

**5.2 Modeling:** We split the dataset into train and test set using random-splitting in the train:test size ratio of 80:20. The training set consisted of 18800 records and the test set consisted of around 4700 records.

**5.2.1 Evaluation Criteria:** We considered the overall accuracy of the model and the F-1 Score as the evaluation criteria for the models. Since, the job role titles are not completely mutually exclusive in terms of job description, we will also look at the top 2 accuracy of the classification model.

**5.2.2 Baseline Model:** We built a DummyClassifier that always predicts the class with the highest empirical probability of occurring. The Dummy classifier has an overall accuracy of 36% with an F-1 Score of 0.19.

**5.2.3 Classifier Candidates:** We tried out Linear/Non-Linear models, Tree-based models with ensembling and Multi-layer perceptrons as classifiers. We train these models on the training dataset, and evaluate on the test dataset.
In order to deal with class-imbalancing we updated the class-weights in the loss function of the different models to account for the under-representation of some of the classes. (Using built-in parameters in sklearn and other libraries)

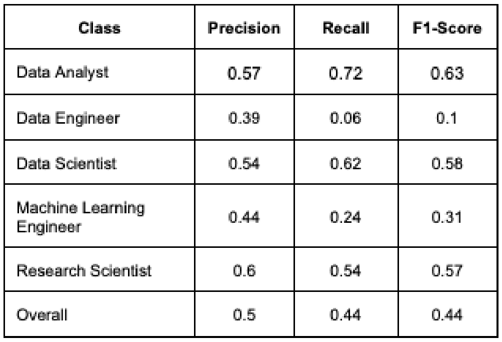
The best performing classifier among all the classifiers is the LightGBM classifier with an accuracy of 55%. We further tuned the hyperparameters of the LightGBM Model to improve the performance.

**Preliminary Results (Classifier Candidates)**

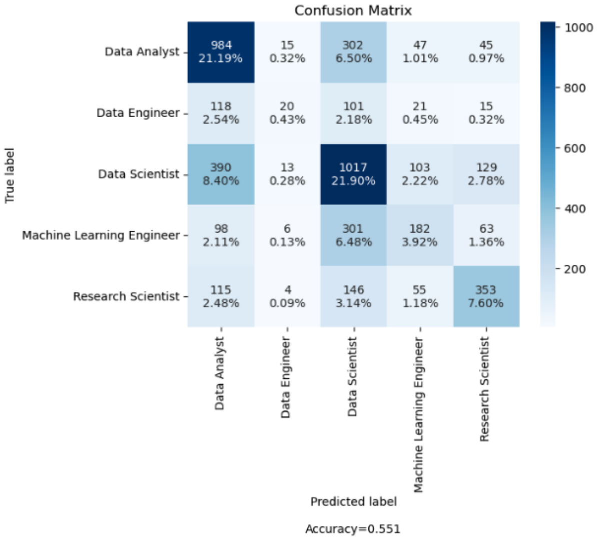
| Model | Accuracy | F1-Score | Top 2 accuracy |
|---|---|---|---|
| Dummy Classifier | 35% | 0.10 | 65% |
| Decision Tree | 41% | 0.41 | 51.9% |
| KNN Classifier | 42% | 0.33 | 62.6% |
| Random Forest | 48% | 0.44 | 69.0% |
| SVM | 47% | 0.45 | 78.6% |
| Multinomial Logisitic Reg | 48% | 0.45 | 71.8% |
| Multi-Layer Perceptron | 54% | 0.43 | 78.8% |
| XGBoost | 54% | 0.45 | 79.1% |
| **LightGBM** | **55%** | **0.45** | **79.23%** |

5.2.4 **Hyperparameter-Tuning:**

i. We tuned learning rate along with hyperparameters that affect the properties of the individual weak learners, such as number of weak learners, maximum-depth of the trees, number of columns to consider while creating a new tree, etc.
ii. We performed 3-fold cross validation by splitting the training set into training and validation sets. The metric that we optimized for here is accuracy.
iii. We performed a Randomized search for the optimal parameters over the pre-defined search space of each of the hyper-parameters.
iv. We got the following best set of hyperparameters: {'subsample': 0.75, 'n_estimators': 600, 'max_depth': 5, 'learning_rate': 0.01, 'colsample_bytree': 0.5}



| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Data Analyst | 0.57 | 0.72 | 0.63 |
| Data Engineer | 0.39 | 0.06 | 0.1 |
| Data Scientist | 0.54 | 0.62 | 0.58 |
| Machine Learning Engineer | 0.44 | 0.24 | 0.31 |
| Research Scientist | 0.6 | 0.54 | 0.57 |
| Overall | 0.5 | 0.44 | 0.44 |



### 5.2.5 **Evaluation:**

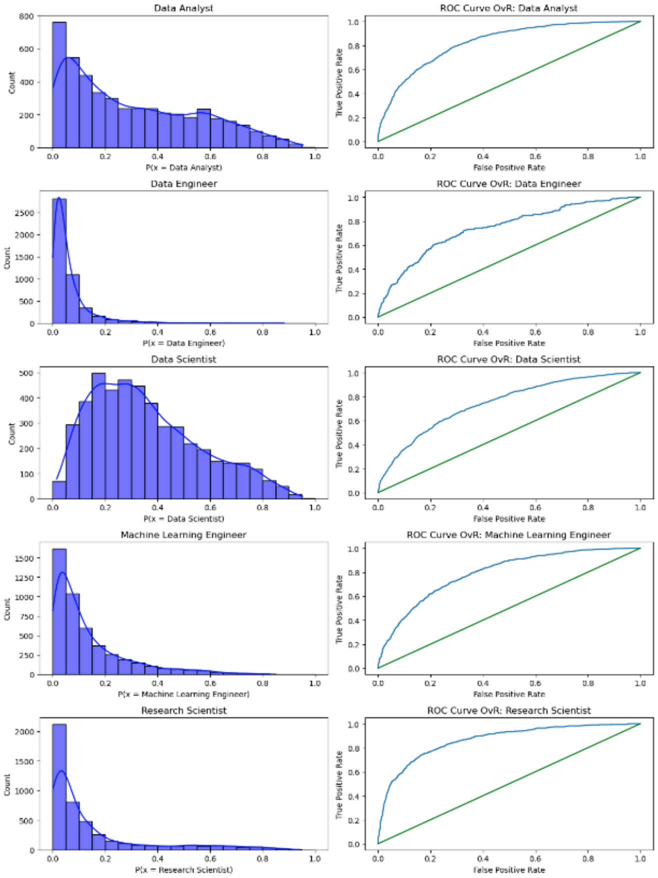
i. The best model has an accuracy of 55.1% and an F1-Score of 0.44, with a top-2 accuracy of 79.62% (figure above)

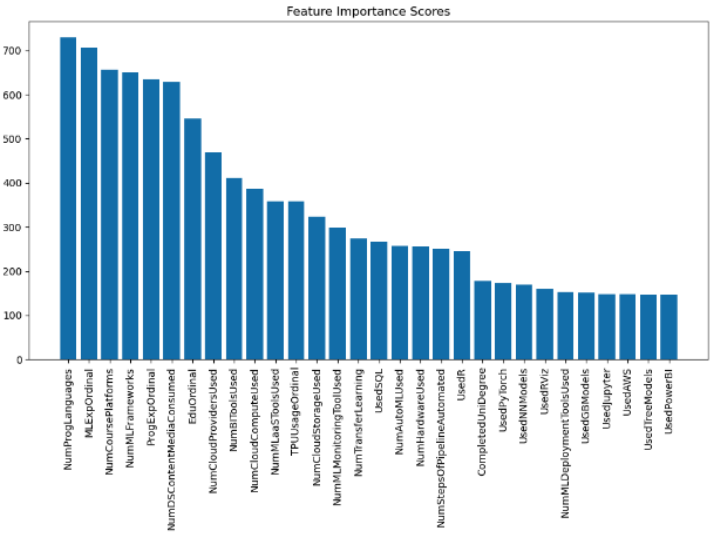
ii. The top-2 accuracy of 79.62% suggests that the model is able to correctly identify the top two most likely classes for a given input with a high degree of accuracy

iii. We have highest F1-Scores for the classes 'Data Analyst', 'Data Scientist' and 'Research Scientist'

iv. The feature importance obtained from the model is plotted (below), and the AUROC curves are plotted for each of the classes (alongside). The F1-Score for Data Engineer is low majorly because of the underrepresentation in the dataset.

v. We have also plotted the distribution of probabilities for each class to judge the confidence of the algorithm regarding it's predictions. The probabilities of Data Scientist class prediction are distributed overall, which implies low confidence of the model generally in predicting the title. For a research scientist the model is highly confident in predicting those who are not Research Scientists





# 6. Conclusion

## 6.1 Summary & Conclusions

We performed four tasks in this project. Firstly, in Hypothesis Testing, we identified that the median salary of Machine Learning Engineers and Data Scientists is higher than that of Research Scientists, Data Analysts or Data Engineers. We also identified that year on year, the proportion of such job roles with advanced degrees is only increasing.

Secondly, in the prediction task, we tried to predict the salary one could demand given their skills in the US. We achieved an $R^2$ of 0.22 on the full dataset and an $R^2$ of 0.40 an adjusted dataset with salaries less than USD 500k. Understandably, the most important features for salary turned out to be years of experience in Machine Learning and Programming. However, we were significantly limited by the lack of predictive features in building a more accurate model. Features like depth of knowledge, the university the practitioner graduated from, etc could have helped make our model more accurate.

In the clustering and classification task, we tried to separate practitioners into different job roles based on their skills and experience. Clustering revealed that the data is not easily separable into job roles, but instead it is separable by the skill and expertise of the practitioners. We concluded that people who have more hands-on experience (more technologies used, more up-to-date with the times) are more likely to demand a higher salary. In the classification task, we experimented with both linear, and non-linear models to classify people into job roles based on their reported skills and experience. The best model achieved a top-1 accuracy of 55.1% and a top-2 accuracy of 79.6%. The F1-score was 0.45. Being a nascent field, the job roles are not easily separable based on skills, and more predictive features like company name and specific role would help improve the accuracy.

### 6.2 Limitations & Assumptions

Our study has numerous limitations. First, the study suffers from self-reporting bias. A respondent may over or under-report their skills and salary. However, we assume that the reported information is truthful. Second, our study is only representative of the people who responded to the survey (self-selection). This is biased towards "Kagglers", and may not accurately represent the whole job market. Third, since data science is a nascent field, the job titles do not have a consistent nomenclature. For example, a Data Scientist in one company may have work responsibilities that are very similar to a ML Engineer in another company. Finally, there is a temporal structure to the data since the responses are collected over a period of three years. However, we ignore it for the purpose of this project. Although, we find that the response year is not an important predictor of salary, the nomenclature of job titles and the skills required for each role may change over time.

### 6.3 Ideal Dataset

The skill-based features in our dataset are binary and only convey if a respondent possesses a skill (breadth of experience), however the level of expertise in each skill (depth of experience) could be an important predictor of both the job title and salary. This would also help us understand what characteristics differentiate individuals with high salaries. Additionally, features such as an individual's prior work experience, the university they graduated from, and their experience with competitive programming could be useful features for salary prediction.

To collect the ideal dataset, we need to collect a representative sample of responses from university surveys, actual companies, and Kaggle. Additionally, we need to refine the questionnaire by adding skill-depth based features and define very specific mappings for each feature response. For example, for the feature "ML Depth", we could define the following categories:

0 "I don't know ML".
1 "I know how ML models such as Linear/Logistic Regression work".
2 "I can mathematically explain the working of the above models".
3 "All the above, and I can code these algorithms from scratch".
4 "All the above, and I have used these algorithms in a real world ML project".
5 "I have programmed/designed advanced ML pipelines, or published novel ML research".

### 6.4 Other Interesting Insights & Future work

Question: Is there a mismatch between the skills of data practitioners and what is listed on the job descriptions?

To answer this question, we scraped over 16k data science Job Descriptions (JDs) from LinkedIn and Glassdoor and used regex and fuzzy matching to extract skills from the JDs. Then we performed a 1-tailed Z-test for proportions (alpha = 0.005) to identify skills that are oversold in job descriptions, but not used on the job, and also skills that are not listed in the job descriptions but frequently used on the job. We found some interesting insights:

#### 6.4.1 "Don't believe the JD, apply anyway!"

○ The proportion of Data Engineers using Java on the job *(20%)* is significantly lower *(z = -5.03, p-value =* $2.49\text{x}10^{-7}$*)* than the proportion of Data Engineer job descriptions that list it as a requirement *(40%)*.

○ Similarly, the proportion of Data Scientists who use R on the job *(43%)* is significantly lower *(z = -5.7, p-value =* $6.05\text{x}\ 10^{-9}$*)* than the proportion of jobs that list it as a requirement *(55%)*.

#### 6.4.2 "Latent Skills you didn't know you needed to learn."

○ Business Analysts should know Python. The proportion of BAs who use Python is significantly greater than the proportion of JDs which list it as requirement *(z = -5.7, p-value =* $6.05\text{x}10^{-9}$*)*.

○ Data Scientists should know ML well. Algorithms like Linear and Logistic Regression, Random Forests and Gradient boosting models are frequently used on the job.

Finally, based on the JDs dataset, we found that SQL, Python and Java are the top skills that Data professionals should have.

## 7. References

[1] “2022 Kaggle Machine Learning & Data Science Survey.” https://kaggle.com/competitions/kaggle-survey-2022 (accessed Dec. 20, 2022).

[2] “2021 Kaggle Machine Learning & Data Science Survey.” https://kaggle.com/competitions/kaggle-survey-2021 (accessed Dec. 20, 2022).

[3] “2020 Kaggle Machine Learning & Data Science Survey.” https://kaggle.com/competitions/kaggle-survey-2020 (accessed Dec. 20, 2022).

[4] “Download all data from our salary survey for free,” *ai-jobs.net*. https://ai-jobs.net/salaries/download/ (accessed Dec. 20, 2022).

[5] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998, doi: 10.1023/A:1009769707641.