Shauna Smith

Bellevue University - DSC680

Professor: Catherine Williams

Week 7 Project2 Milestone3 "Data Dream Jobs Model Final Copy"

Project 2 - "Drata Dream Jobs"

History/Buisness Problem

```
Project Topic - "Predicting Job Type by the select desireability of life-style
expectations to best serve Your university Major and ultimate career goals."
```

At Kukis-Garo we are the country's leading inovator, acting as the top headhunter for Universities and the like. We have a long history of dealing with hunting heads, and in the modern world we are just as excellent at hunting. Kukis-Garo works hard to find the greatest advantages for implenmenting both efficient and provable metrics to aquire the greatest minds from the sought out greatest heads to apply towards any goal.

For Universities, we pride ourselves in finding great students and developing their drive towards a career based goal education. We desire to see fullfillments from advanced learning all the way into career fields of their choice. The proposal at hand seeks to deliver just such an initiative and provide a method for predicting the job types as a motivationaly based expectation of comitments from the student's preferences. To best do this, we are utilizing a means of predictive analytics & employing ML, to better offer comprable real-world data that predicts what you need for success. Ultimately, this data collection and model will yield predictions that better suite the needs for career types. It is derived from the selective input expectations customized for each student and sourced for each teacher or program offering. In turn, this should help maintain enrollment retentions to better improve graduation rates on the whole. As a collective assesment process, it will implement metrics of improvements that will encourage future student prospects and beneficially increase the reputational outlooks of each University that endorses such methods.

Data Explanation(Data Prep/Data Dictionary/Etc.) & Methods

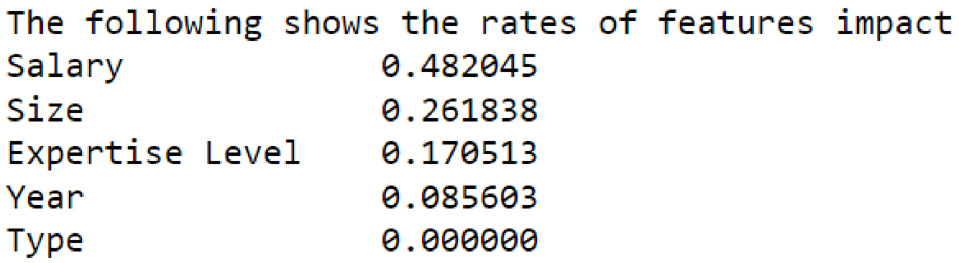
The Dataset used for this framework was derived from the open-source Kaggle found at the following link: https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023

This dataset consisted of Data Science related Job types and their corresponding features. Some of the features used were that of Salary, Employment Type, Expertise Level, and Job Titles. All these features are relegated from a relevent given time span of 2020-2023 and thus appropriately allow for a real-world framework as conscripted to current estimations. This practice Model will serve to catagorize data related fields of education and directly hone in on servicing future Data Scientist related programs or other such top data technology related studies.

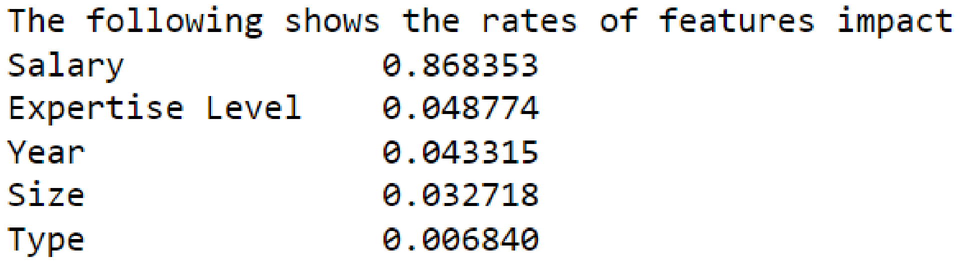
For preperation purposes, we started with a relatively clean dataset that had no missing values and we chose to drop the features that were redundant or unuseable for our intended pupose. This led to the elimination of some features such as country based metrics for varying types of currency and the like.

Then, by assigning renamed deliniation methods to the values in the working features we formed a better machine suited dataset. This consisted of taking the unique values and subgrouping them into qualities represented from the whole. Such an example of this was regarding Expertise Level, in where it was broken down into numerical replacements of 1 or 2(Higher expertise vs lower expertise).This assignment was derived from an original starting measurement of 'Expert','Director', 'Intermediate', 'Junior' from which the original data observations accounted for. Similar steps were taken to Company size or others, but some of these subsequently were shown as feature with lacking importance, and therefore they were dropped from the final working dataframe in the end.

Feature Importance from DTC and RF



```
The following shows the rates of features impact
Salary             0.482045
Size               0.261838
Expertise Level    0.170513
Year               0.085603
Type               0.000000
```



```
The following shows the rates of features impact
Salary             0.868353
Expertise Level    0.048774
Year               0.043315
Size               0.032718
Type               0.006840
```

Data Dictionary:

Feature	type	# used +/-
Job Title	**string**	**as+"Job"**
Employment Type	**string**	**-(redundant)**
Experience Level	**string**	**-(redundant)**
Expertise Level	**integer**	**as+"Exper..."**
Salary in USD	**integer**	**as+"Salary"**
Company Size	**string**	**-(redundant)**
Year	**integer**	**as+"year"**

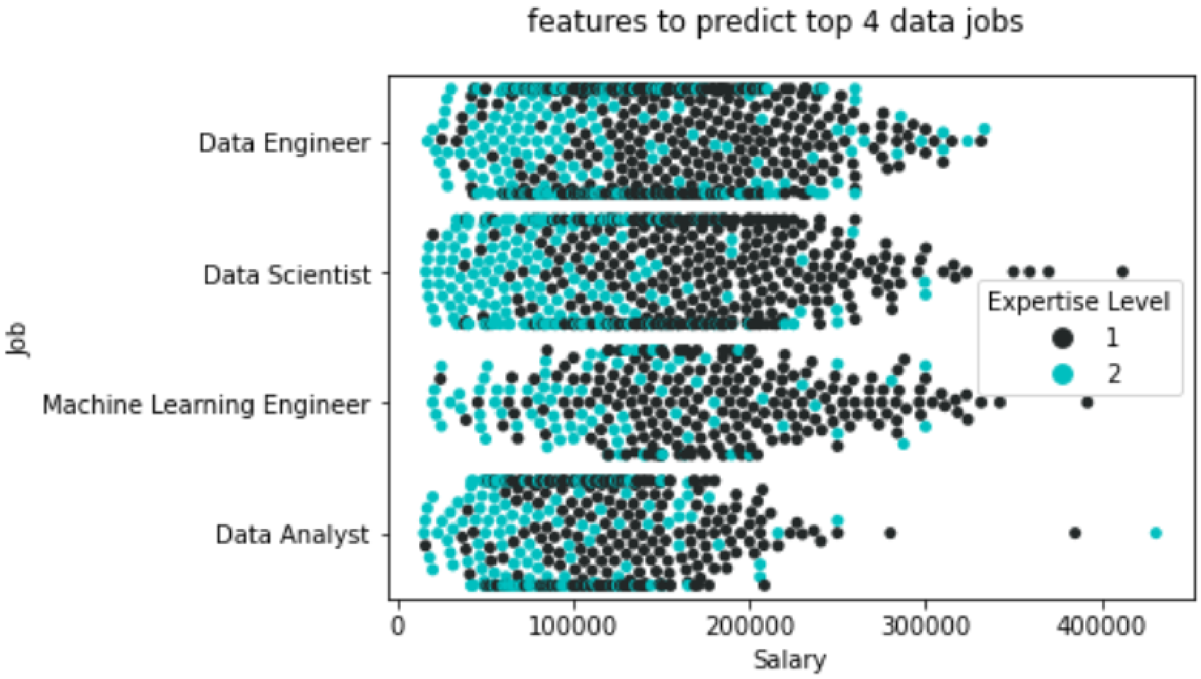
Several data prepping approaches were employed to the dataset during the discovery process and steps to better adjust the dataset features to work with the elements related to the model's preferences were applied as needed to best develop the output. The application for predictive analysis using Random forest and decision Tree classifiers were employed. This gave a better insight into the feature importance. Final additions for developing a

linear regression analysis aided in building useable visualizations but yielded extremely low accuracy predictions as a model construct.

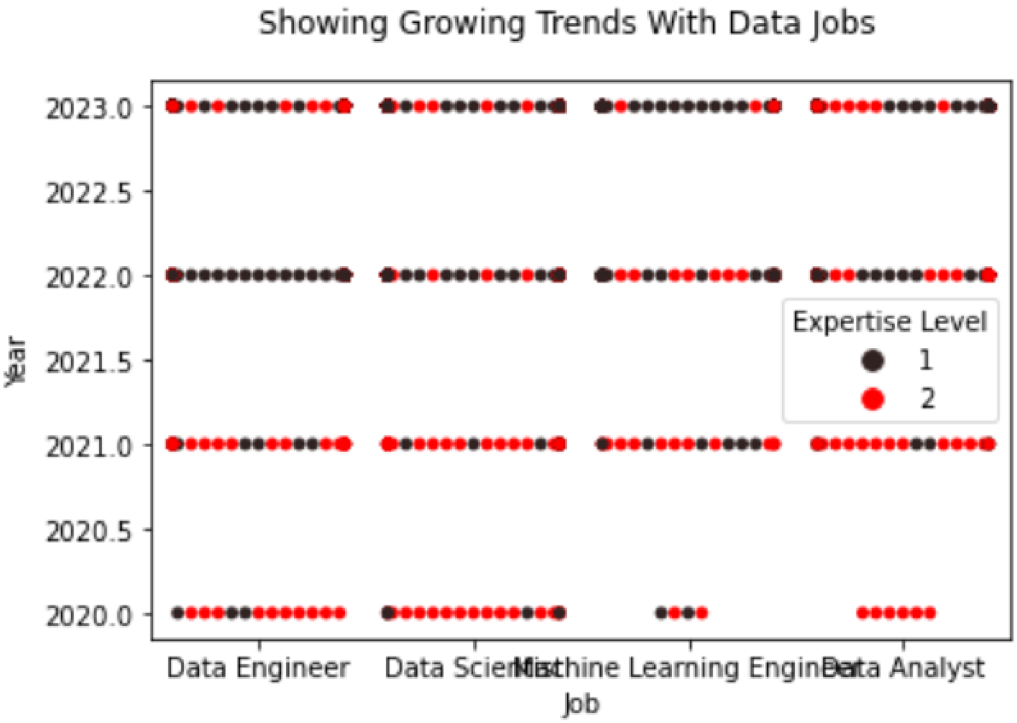
Analysis/Conclusions With Assigned Assumptions & Limitations Or Challenges

In the end, the best performance was gained from the DTC(Decision Tree Classifier) and even though the accuracy score was very low at approxemately only 22%, it does repeatably demonstrate the feature importances for a good predictive nature inherent to the dataset. This demonstration dictates "Salary" as the highest predictor, followed by "Expertise Levels" and ultimately gives us a starting place as an acting framework to build on.

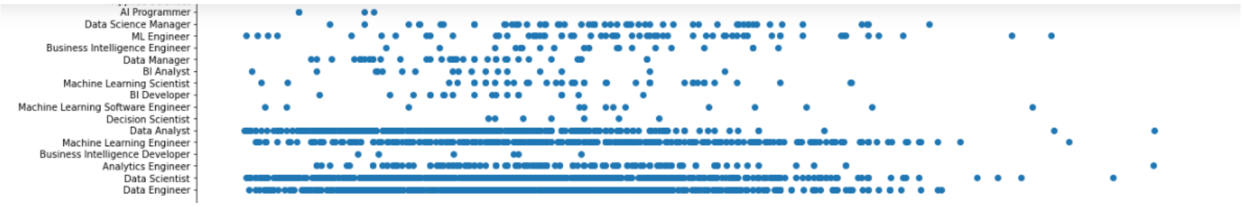
Unfortunately there are very clear limitations to the existing datapoint observations, but this does not entail an inability to become a workable strategy as shown in the visualizations. With the observations given, the aspects related to a "Machine Learning Engineer" as an example, it shows how we can easily predict a large trend represented for a given timespan that increases in interest as expertise. Proving this is a worthy Job title in the dataset for further study as expertise levels develop even further with time:



And here again, we can see how all of the top 4 data Job titles will gain in their expertise level over time, which shows a trend for a growing intrest and demand in the field of data jobs:



Although the largest challenge is the low accuracy score as related to the nature of the current working dataset, it is believed that as growth in the observations develop over time, so too will the accuracy score from better training data. Therefore, this endeavor is still one of worthy pursuit, but will require further investments or subsidies before it can provide an admirable accuracy score duplicated by the ML model itself. I would advise assisting in the efficient development of the model through the means of more observations and a decided focus for select Job titles from the open field. The following dataframe selects from the highest occuring titles and serves as the proposal for limiting job types to the top 4 given data jobs and respective course programs through universities:



<u>Future Uses Or Additional Applications, Recommendations, & Implementations With Ethical Assessments</u>

There are several available methods worthy of employing the predictive functions for our model. We could attribute a DTC(Decision Tree Classifier), an ensemble RF(Random Forrest), or even define a multivariate Liner Regression model once adequate collections amass as a worthy set of observations. However, the goal will be to incorporate features that best atribute to a classification prediction for a "Target Job" as the elected Job Types found in the "Data Science Salaries" dataset from Kaggle. Further emphasis on limiting or complimenting desired course programs should also be considered when applying supplemented observations from which to train.

As for Challenges, finding the best method to produce the highest accuracy for predictions will be limited by the available datapoint observations until time has amassed a broader working dataset, and the selective features derived from the given dataset should be maintained. However, since this is intended as a framework at current, any additional future subsidy will also escelate the findings to train for a more robust model, but also serve as research and improvement metrics during the post and collective interim. The biggest issue, will be on how to implement that framework and moderate the functions thereof in an efficient and expedient manner that does not interfere with student motivations.

As of now, there are no known ethical considerations as it is entirely a subjective modus in way of preferences or performance by the given user, and establishes a real-world evolving assesment to best assume the predictive nature of the provided Job types. Assigning these career goals only serves as a start to finish customization of educational goals based on the predictive quality of the output. Therefore, they are intended as a guidence, but not guarentee. Since there are no guarantees, nor assesment metrics presented on behalf of the users, there are also no permanent nor demonstrative markers issued into the identity of any given users. Thus, due to this subjective preference and best current assignment for correlative values, no ethical contstraints are necessarily considered nor needed.

Questions we should ask:

- **How is this going to help students? This model seeks to help students by showing achievable real-world goals as the proverbial trophy from their hunt.**
- **How is this going to help Universities? This model will help Universities by gaining not only feedback as implemented metrics of content, but also sourced knowledge as research and cirriculum comprehension.**
- **What are the expected costs associated with maintenance for this model? This model will aid in supplemental income and content, but maintanence upkeep, and storage needs are outsourced under the most efficient means.**
- **Should we concern ourselves with regulatory measures? As the aspects of data content are relegated to non-identifying subjective basis, or real-world figures derived from current and open sourced metrics no concerns regarding regulatory constraints are currently needed for evaluation purposes.**
- **Could we issue a faster method for collecting, subsidizing, or improving any gains yielded from the proposed model? As innovation is ever changing speed is usually a measuremnt for improvement, but as of now the proposed model gains a fast insight into the predictive qualities assigned by the data inputs provided. As those datasets grow over time, so too will innovative applications into yielding their outcomes.**
- **Are there methods that supplement or replace this endeavor of discovery? Much of the predictive qualities derived from this model are readily available and of interest often found in current event editorials. However, these are limited to a generalized assay, and scope. Where our model differs is in application of the predictive nature customized for each inquiry and assesment purposes.**
- **How beneficial or well received will this implemented step for students be? It is with high hopes that the process of incorporating survey steps under the guise of improving outcomes tailored for each individual will be met with optamistic endorsements. Additional substraints that impress on both the validity and reputational dev3elopment of Universities that employ these measures will be perceived as highly motivated in the partnership for sucess.**
- **What are some open sources we could use to further supplement observations, or is other data augmentation methods available to us? Although data augmentation is an option with model selections, it is not recommended. The goal of the model is to be acurately representative of the current demands and growth trends regarding the given subject matter of these specific job types. Therefore, real-world sourcing is the best method for filtering subsequent data prior to adding it as a working observation. By keeping this strict standard and incorporating collective student survey follow ups, we help train a more robust model and provisional reputation for guidence on these matters.**
- **How long is expected before yielding outputs worth mitigation? It is believed that outputs are useable as motivators as current, the predictive accuracy of the model will improve with time as observations derived from mitigation are developed. The direct worth of mitigation stems from the show of care and tailoring to the students needs and preferences towards a shared goal of program completion. Thus, mitigation alone makes the effort a worthwhile endeavor.**
- **Are goal oriented motivations the only expected benefit from having a ML model of this type? Implementation of this model will reap more than just motivational charms. Simply by applying these measures, improvements for quality metrics are enforced, and student centered success is highlighted. However, reputational payout will service not only direct costs, but gain a return on the investment at exponentially grand rates via prospective and quality proveability.Simply by way of endorsement, many rights of integrity and research sourcing means can function as supplemental to the model's designed goals, but ultimately it will yield far greater profitability by way of processing customized dependencies into the future.**

All validation of output accuracy will be gained from the chosen method of the model implemented. As for the outcomes, follow up surveys or retention rate increases of the users within the pool of model participation are excellent means for performance metrics. Having the comparative analysis of projected accumulations accounting for before the model's use vs after, could even prove as productive or inconclusive. Analyzing other open sourse datasets, or available updates that compliment existing feature markers are another means of supplementing observations, but guidence to relevance should be cautioned in the development.

However, imploring techniques as a motivator for student completion, is a valid trade off for headhunting students to better fit for their needs. This will aid in graduation rates and subsequent reputation building efforts. Also, issuing survey rounds to analyze user content towards efforts employed during and after program attendance, could provide a means for additional improvement metrics or advice.

As a final note, this survey proposition has no bering on performance, but could be desired for additional marketing or extending research purposes. At Kukis-Garo, we are always looking forward to better tomorrows where dreams await us, and cooler heads prevail.

Appendix

Definitions

TERM	MEANING
RF	**(Random Forest classifier)**
DTC	**(TDecision Tree classifier)**
Size	**(Company Size as large, medium, small)**
Job	**(Job Titles)**

Reference

BANERJEE,SOURAV(2023) "Data Science Salaries", Kaggle.com, https://www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023

```
In [ ]:
```