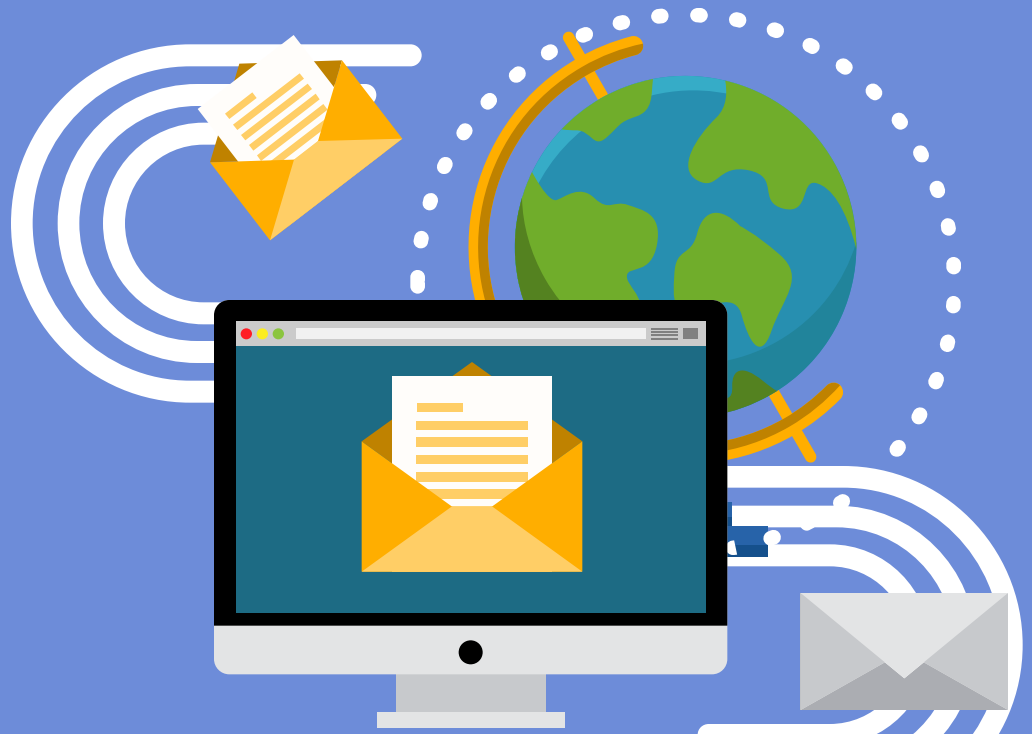
Regression on Annual Compensation

Dataset from 2018 Kaggle ML&DS Survey Challenge

Yifan Zhang
1004245952

This page is not included in 5 slides.



Data Cleaning and Data Analysis

	Unnamed: 0	Time from Start to Finish (seconds)	Q1	Q1_OTHER_TEXT	Q2	Q3	...	Q50_Part_5	Q50_Part_6	Q50_Part_7	Q50_Part_8	Q50_OTHER_TEXT	index
0	2	434	Male	-1	30-34	Indonesia	...	NaN	NaN	NaN	NaN	-1	0.0
1	3	718	Female	-1	30-34	United States of America	...	NaN	NaN	NaN	NaN	-1	1.0
2	5	731	Male	-1	22-24	India	...	Not enough incentives to share my work	NaN	NaN	NaN	-1	2.0
3	7	959	Male	-1	35-39	Chile	...	NaN	I had never considered making my work easier f...	NaN	NaN	-1	3.0
4	8	1758	Male	-1	18-21	India	...	Not enough incentives to share my work	NaN	NaN	NaN	-1	4.0

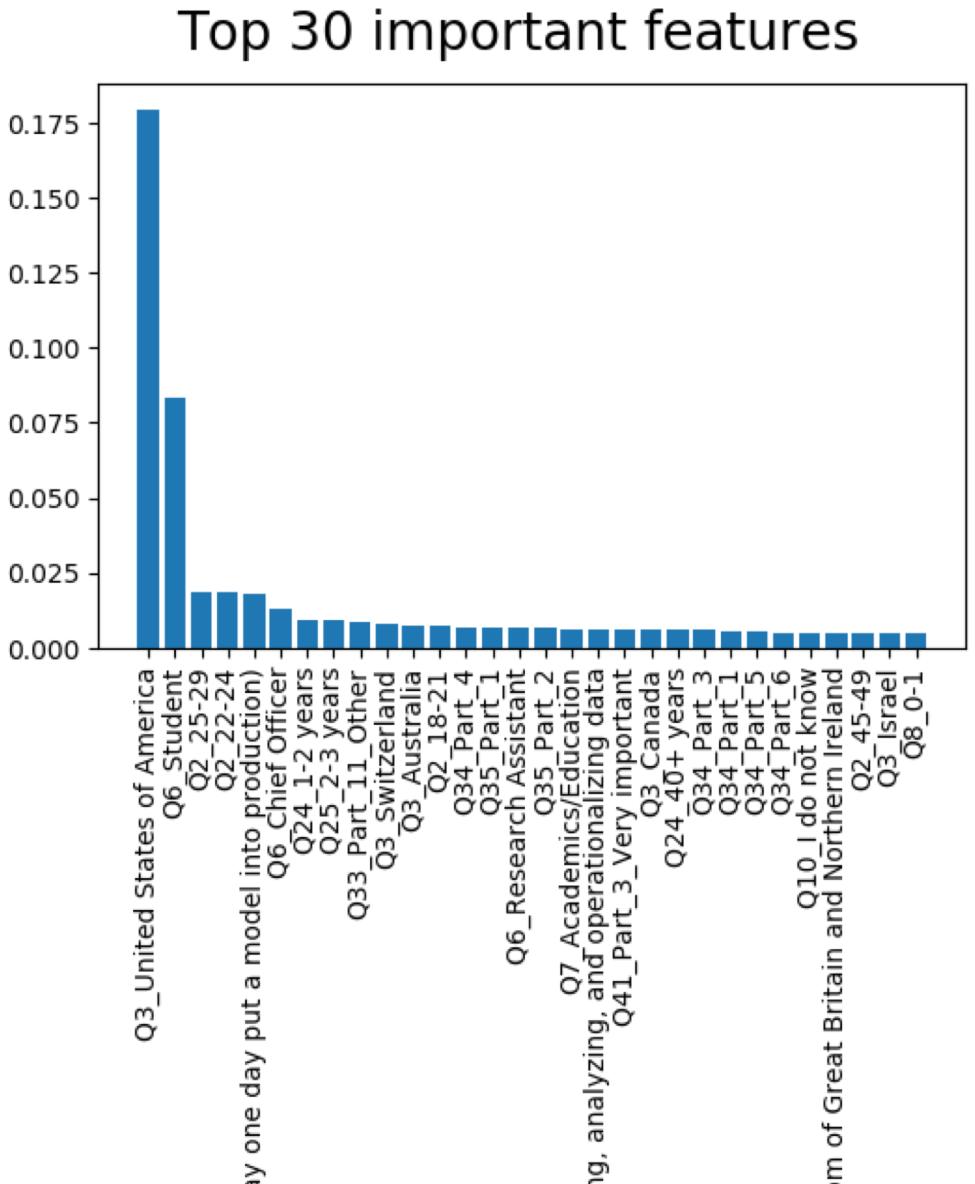
- As for numerical features:
  - titles of features contain "OTHER_TEXT" are dropped
  - other numerical features are standardized
- As for categorical features,
  - create new column for each category and rename the title of the column to the former name appended with the category name.

- Drop the feature 'Time from Start to Finish (seconds)'
- There is almost no valid answer in feature 'Q38_Part_19' and 'Q38_Part_20', so we decide to drop those two columns
- Drop the rows with sufficient information.

From 397 features to 655 features

	Q1_Female	Q1_Male	Q1_Prefer not to say	Q1_Prefer to self-describe	Q2_18-21	Q2_22-24	...	Q50_Part_3_Requires too much technical knowledge	Q50_Part_4_Afraid that others will use my work without giving proper credit	Q50_Part_5_Not enough incentives to share my work	Q50_Part_6_I had never considered making my work easier for others to reproduce	Q50_Part_7_None of these reasons apply to me
1	1	0	0	0	0	0	...	0	0	0	0	0
2	0	1	0	0	0	1	...	0	0	1	0	0
3	0	1	0	0	0	0	...	0	0	0	1	0
4	0	1	0	0	1	0	...	0	0	1	0	0
5	0	1	0	0	0	0	...	0	1	0	0	0

Model Feature Importance

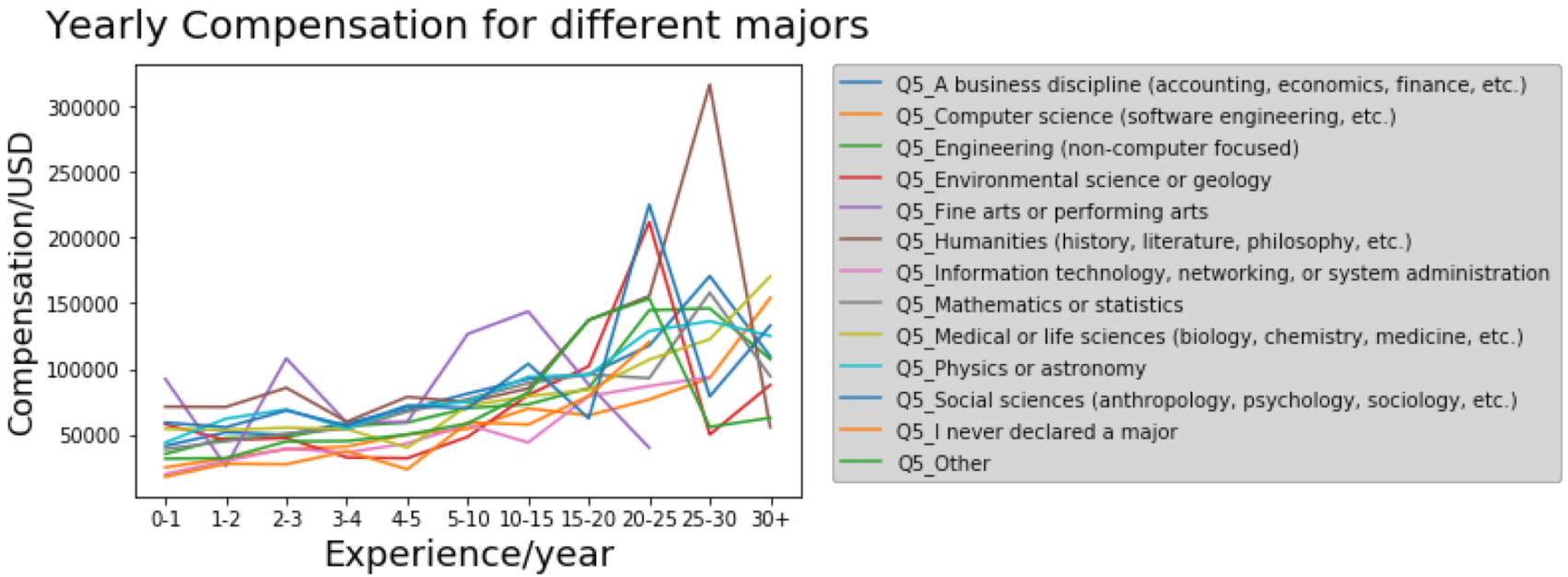


	Q3_United States of America	Q6_Student	Q2_25-29	Q2_22-24	Q10_We are exploring ML methods (and may one day put a model into production)	Q6_Chief Officer	Q24_1-2 years	Q25_2-3 years	Q33_Part_11_Other	Q3_Switzerland	Q9
Q3_United States of America	1	-0.037	-0.019	-0.055	-0.045	0.0086	-0.07	0.032	0.02	-0.048	0.42
Q6_Student	-0.037	1	-0.057	0.2	-0.018	-0.053	0.084	-0.069	-0.0098	0.0014	-0.28
Q2_25-29	-0.019	-0.057	1	-0.29	-0.014	-0.051	0.056	0.067	-0.0019	-0.0095	-0.12
Q2_22-24	-0.055	0.2	-0.29	1	0.024	-0.053	0.15	-0.023	-0.013	-0.015	-0.22
Q10_We are exploring ML methods (and may one day put a model into production)	-0.045	-0.018	-0.014	0.024	1	0.017	0.048	-0.011	-0.02	-0.0069	-0.034
Q6_Chief Officer	0.0086	-0.053	-0.051	-0.053	0.017	1	-0.034	-0.00055	0.0092	0.0044	0.16
Q24_1-2 years	-0.07	0.084	0.056	0.15	0.048	-0.034	1	-0.06	0.0011	-0.0073	-0.17
Q25_2-3 years	0.032	-0.069	0.067	-0.023	-0.011	-0.00055	-0.06	1	0.004	-0.0029	0.061
Q33_Part_11_Other	0.02	-0.0098	-0.0019	-0.013	-0.02	0.0092	0.0011	0.004	1	0.016	0.0096
Q3_Switzerland	-0.048	0.0014	-0.0095	-0.015	-0.0069	0.0044	-0.0073	-0.0029	0.016	1	0.073
Q9	0.42	-0.28	-0.12	-0.22	-0.034	0.16	-0.17	0.061	0.0096	0.073	1

- Visualize the top 30 important features as following in the left graph, there is no doubt that feature Q3_United_of_America, which mean the participant's region is US, is the most important feature for yearly compensation.

- Feature **[region is American]** has strong positive relationship with compensation, while feature **[student] [age: 25-29] [age: 22-24]** has a negative relationship with compensation.

Visualization



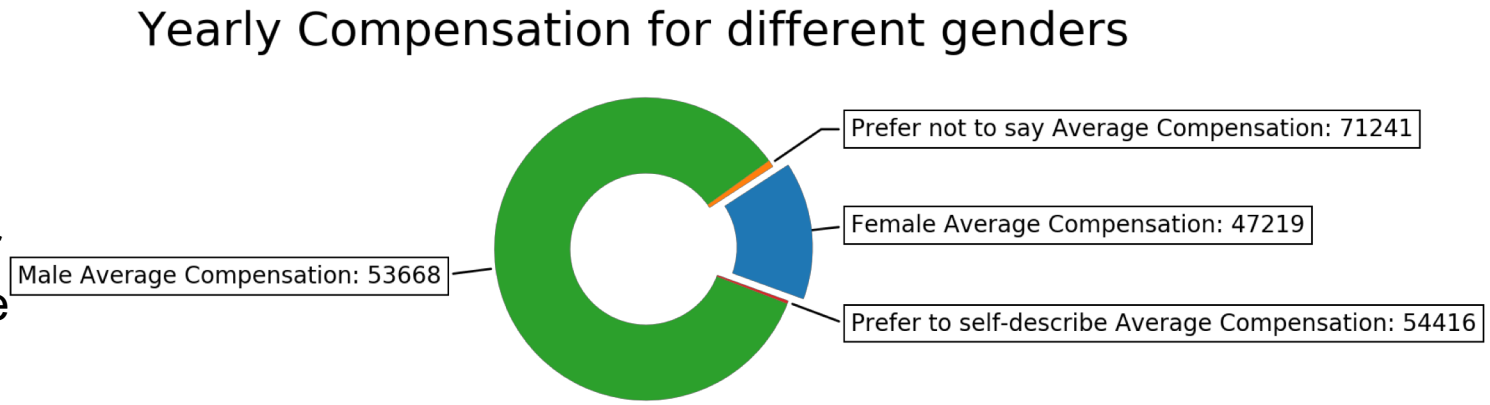
- With the increase of work experience, the annual salary is gradually increasing in general.
- But after 20 to 30 years of work, there has been a sudden turn in the trend of annual salary.
- This may be because of retirement or from a high-intensity job to a relatively comfortable job because of age.

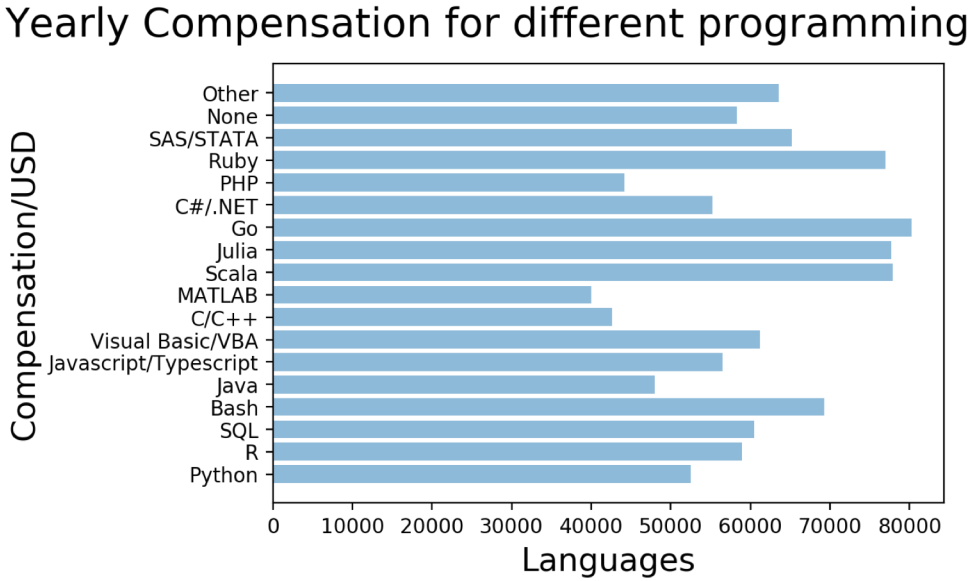
- Humanities shows a remarkable surge during 25-30 and drop down immediately. It might due to some special case, but we should not regard those majors as relatively low income.

- Another finding is that computer science and statistic are not that remarkable among all majors.

Visualization

- The average compensation of male is 53668, which is over 13 percent higher than female.
- People prefer not to reveal their gender or want to self-defined have remarkable average income.





- The most frequently used languages like JAVA, Python, R, C/C++ and JavaScript doesn't show a good performance according to this survey.
- In contrast, Go, Julia, Scala and Ruby, which are kind of new to the field, show higher return.
- The reason is that those new programming languages are powerful in specified fields and have a great need for employees, while only few people master those skills and meet the market demand.

Model Result

Model	Train accuracy (Kfold, R2)	Validation accuracy (Kfold,R2)	Optimal Accuracy (Grid search)	Optimal Parameters (Grid search)	comments
Linear Regression	0.45	0.43			quick
Random Forest Regressor	0.89	0.27	0.276	'min_samples_split': 2, 'n_estimators': 100	Overfitting and slow
Gradient Boosting Regressor	0.55	0.31	0.36	'learning_rate': 0.25, 'loss': 'huber', 'max_depth': 5, 'min_samples_split': 1000, 'n_estimators': 100	Overfitting, Tuning hyperparameter could be helpful
LASSO	0.45	0.43	0.43	'alpha': 1	Quick and fit the model, Optimal solution

LASSO acurracy on testing set is 0.44.

The model's performance is consistent on both training and testing, which nor describe the random error in training set or insufficiently describe the data.