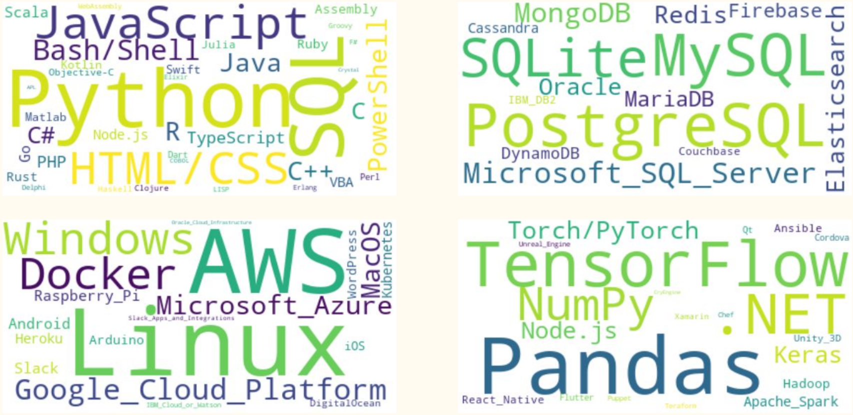
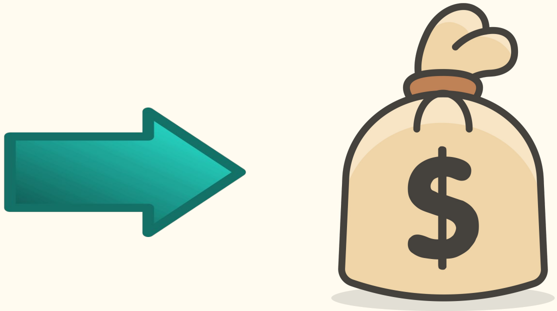
PayUP, maybe?

Charlie Boatwright, Mai La, Matt Pribadi, Jacquie Nesbitt

Motivation

Create machine learning model that helps data scientists identify a salary range for salary negotiations for different opportunities and skill sets



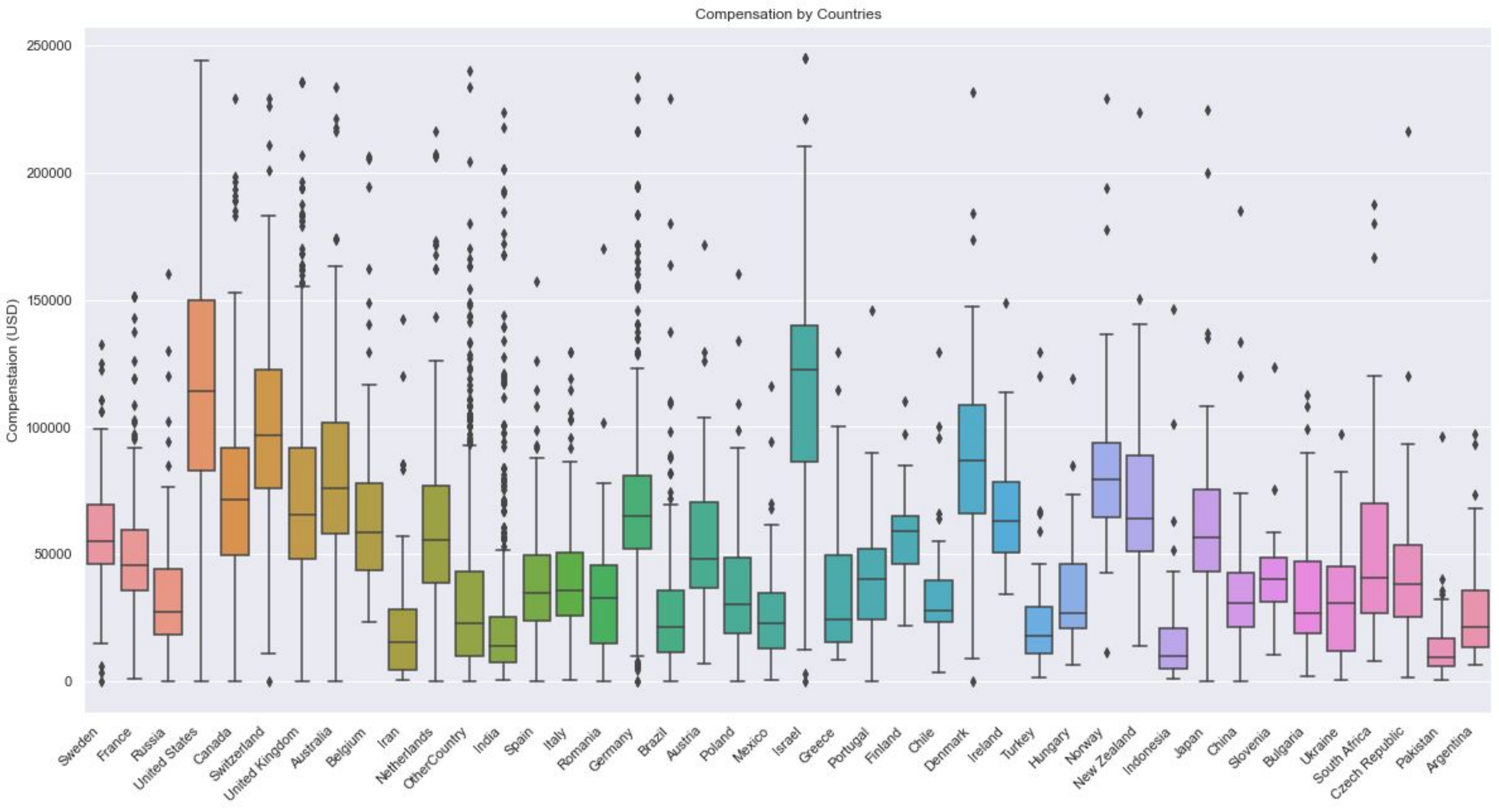


Data

Data Source	Stack Overflow Developer Survey Years: 2019, 2020, 2021
Sample Size	8,331 (cleaned, <$250K)
# of Features	75
Continuous Outcomes	1 (Compensation in USD)
Categorical Outcomes	2 (Compensation Bracket and HML)

```
Train Data Dimension: (5831, 75)
Train Label Dimension: (5831, 3)
Development Data Dimension: (1250, 75)
Development Label Dimension: (1250, 3)
Test Data Dimension: (1250, 75)
Test Label Dimension: (1250, 3)
```

Target - Distribution



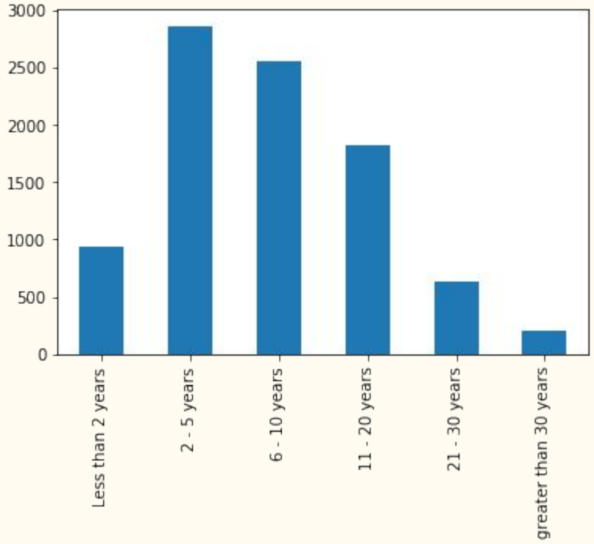
Data - Main Features and Summary Statistics

- Main features can be categorized as:
    - Skillset
    - Individual and Employer Characteristics

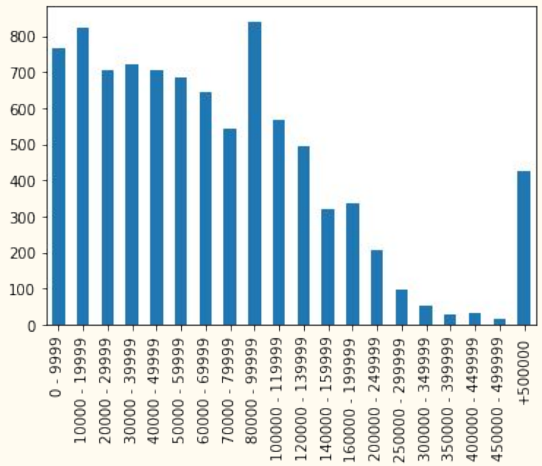
Country

Country	
United States	2316
India	679
Germany	639
United Kingdom	582
Canada	322

Years Coding Professionally

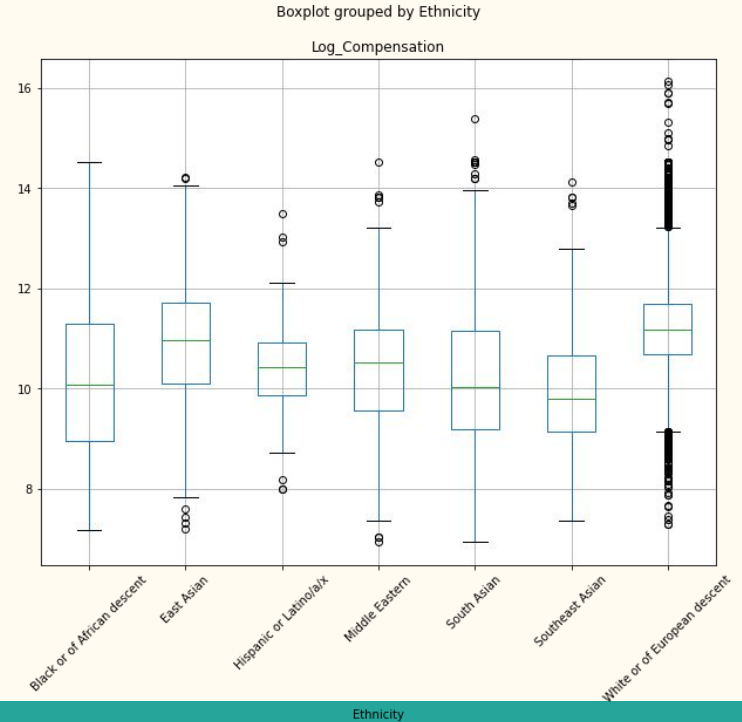


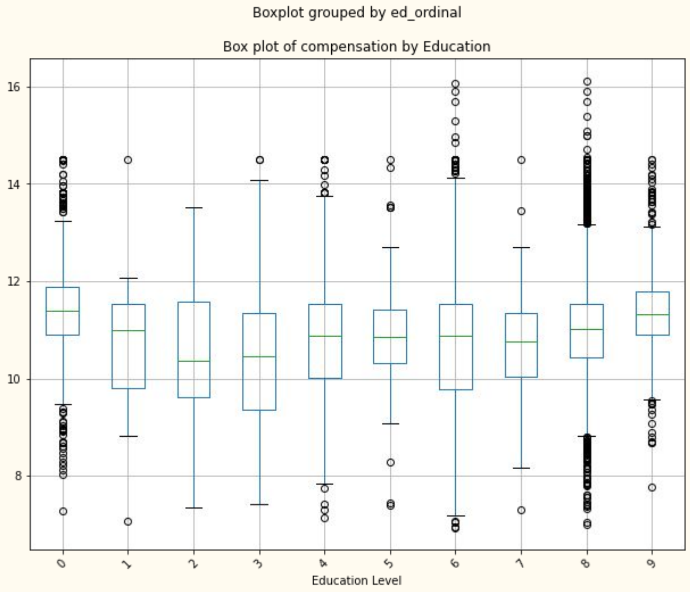
Income Bracket (USD)



Data - Main Features and Summary Statistics

- There appears to be a relationship between ethnicity and compensation.
- Education level and compensation have a slight relationship.

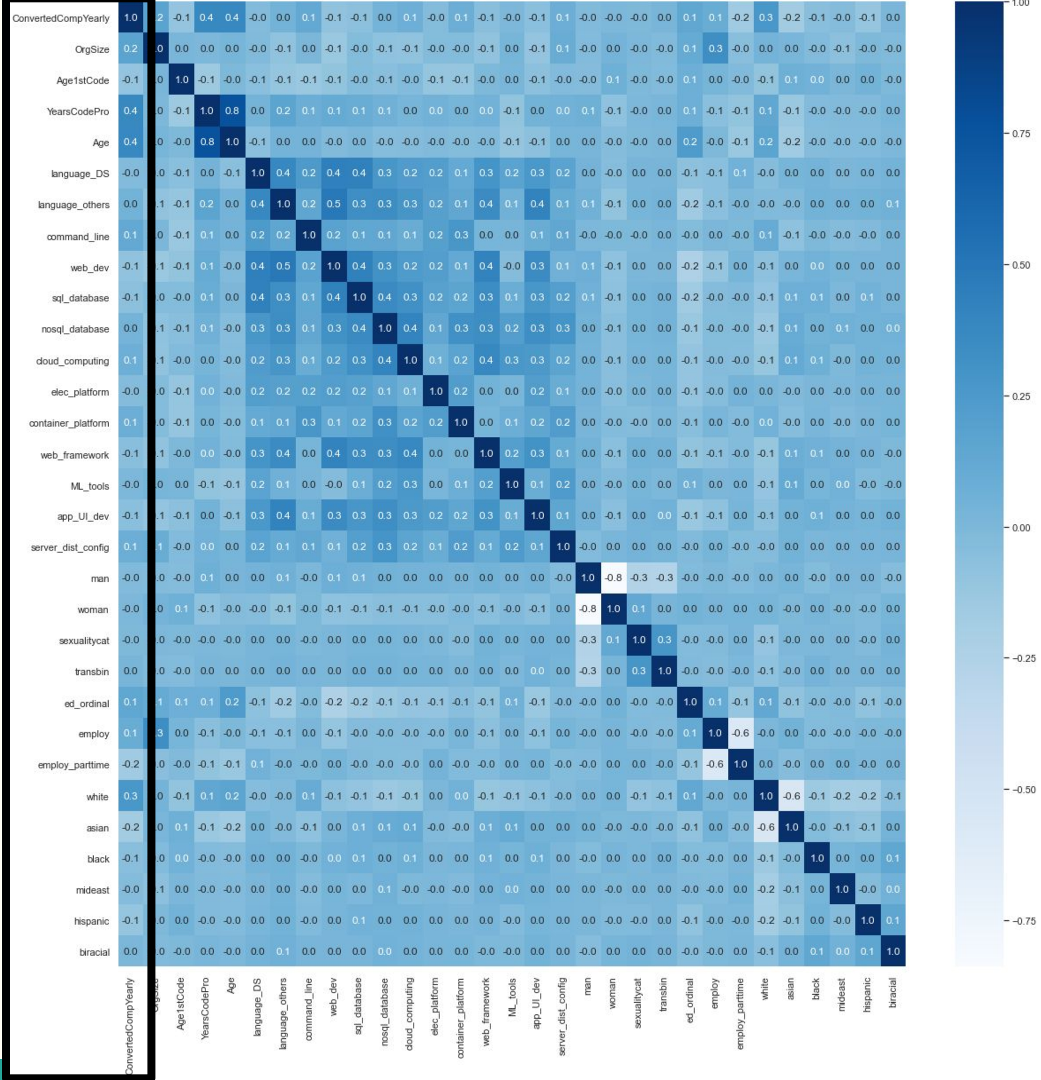




EDA

Low correlation - target & features:

- Best 0.4: years of professional coding & age
- Next best 0.3: white demographic
- Mostly < 0.2



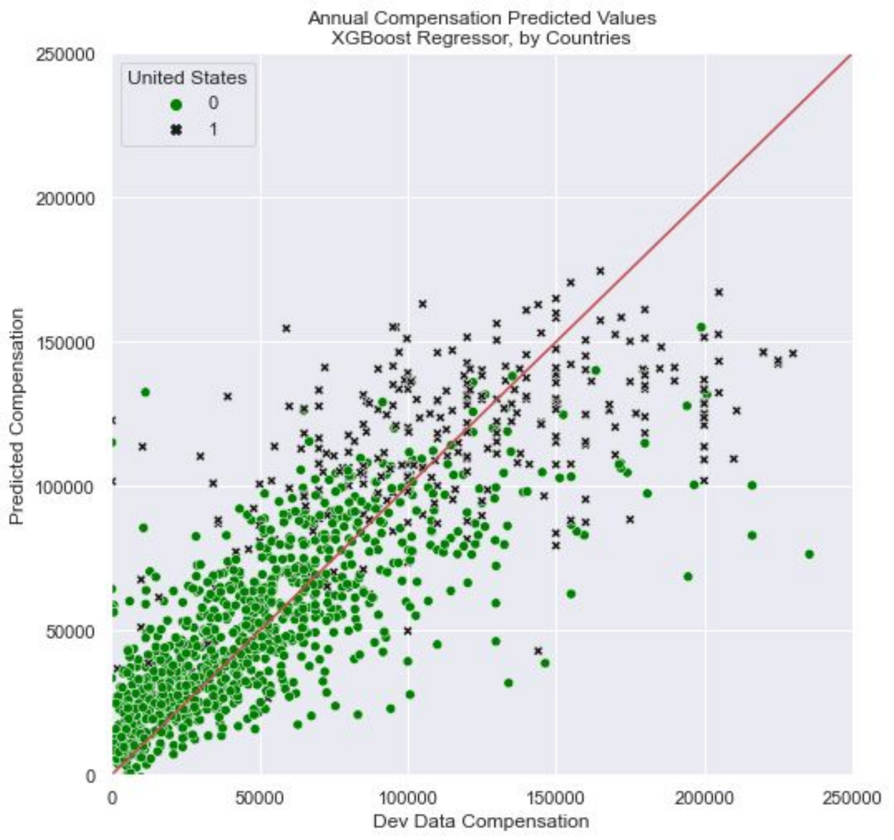
Winning Model

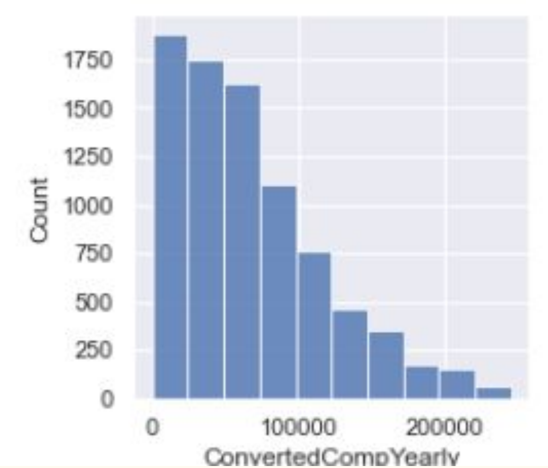
Best Model - XGBoost Regressor

Model	RMSE (USD)	R-Squared
Best - XGBoost	29,166.97	0.656
Base - Regression	30,144.14	0.632

```
Best max_depth: 10
Best colsample_bytree: 0.5
Best L2 regularization: 100
Best n_estimators: 150
Best learning_rate: 0.1
```

Winning Model

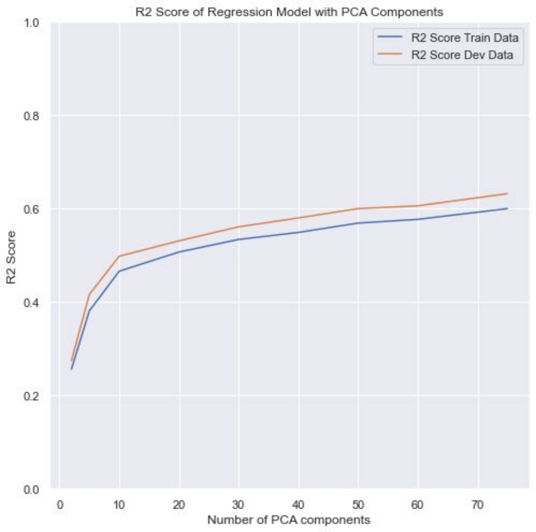


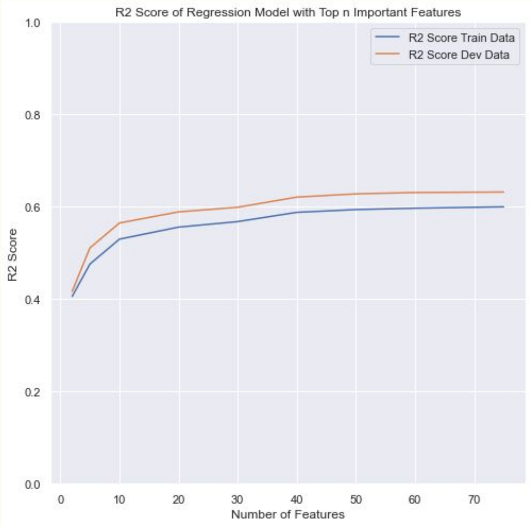


Feature Extraction & Feature Selection for Regression

- Feature selection using Random Forest performs better than feature extraction with PCA
- Using feature extraction or feature selection does not help our model

Model	RMSE	R-Squared
Base - Regression	30,144.14	0.632
Regression & PCA 50	31,449.84	0.600
XGBoost & PCA 50	30,720.91	0.618
Regression with 50 most important features	30,311.20	0.628





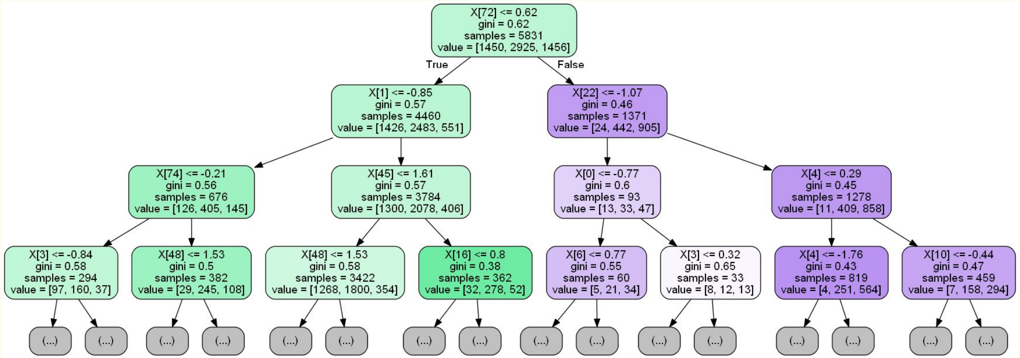
Additional Regression Models

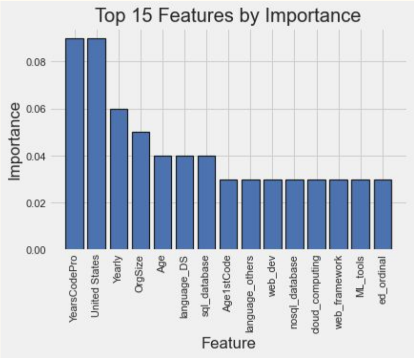
XGBoost > GradientBoost > RandomForest > AdaBoost > OLS/ Ridge/ Lasso > SVR

Model	RMSE (USD)	R-Squared	Hyperparameters
Best - XGBoost Regressor	29,166.97	0.656	max_depth=10, n_estimators=150, colsample_bytree=0.5, lambda=100, learning_rate=0.1
Gradient Boosting Regressor	29,683.04	0.644	max_depth=3, n_estimators=150, min_samples_split=20, min_samples_leaf=5
Random Forest Regressor	29,907.63	0.638	max_depth=30, n_estimators=150, min_samples_split=30, min_samples_leaf=3
ADA Boosting Regressor	29,986.60	0.636	max_depth=30, n_estimators=150, min_samples_split=20, min_samples_leaf=3
Base - OLS Regression/ Ridge/ Lasso	30,144.14	0.632	L1 alpha=10, L2 alpha=2
Support Vector Regressor	30,439.58	0.625	kernel=linear, C=100, epsilon=0.001

Additional Models - Random Forest Classifier

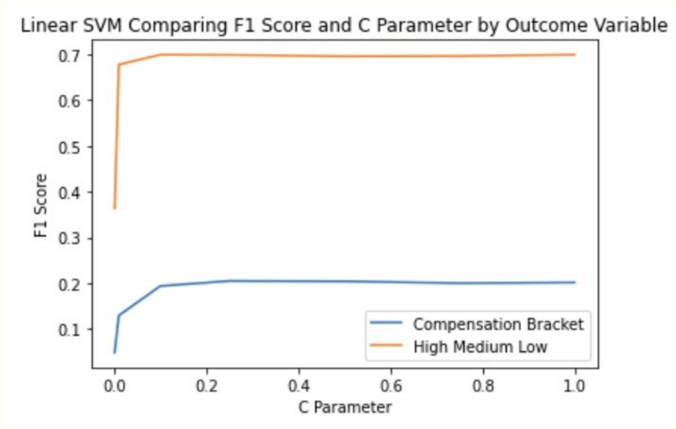
Model	Compensation Bracket (F1 score)	High, Medium, Low (F1 score)
Base	0.242 default tuning	0.727 default tuning
Random Search	0.237 RandomForestClassifier(max_depth=30, max_features='sqrt', min_samples_split=5, n_estimators=500)	0.731 RandomForestClassifier(bootstrap=False, max_depth=30, max_features='sqrt', min_samples_split=10, n_estimators=1788)
Grid Search	0.261 RandomForestClassifier(bootstrap=False, max_depth=40, min_samples_split=10, n_estimators=400)	**0.732** RandomForestClassifier(bootstrap=False, max_depth=40, max_features='sqrt', min_samples_split=10, n_estimators=1750)





Additional Models - SVM Comparison

Model	Compensation Bracket F1 Score	High, Medium, Low F1 Score
Linear Baseline	.189 C = 1	.693 C = 1
Linear	**.226** C = 10	0.739 C = 1



Additional Models - SVM Comparison

Model	Compensation Bracket F1 Score	High, Medium, Low F1 Score
RBF Baseline	.16 C = 1.0, Gamma = .005	.70 C = 1.0, Gamma = .005
RBF	.219 C = 100, Gamma = .005	.735 C = 10, Gamma = .005

Gamma:

- Scale: 1 / (n_features * X.var())
- Auto: 1 / n_features

Additional Models - Logistic Regression

Model	Compensation Bracket F1 Score	High, Medium, Low F1 Score
Global	0.246	0.743
US	0.153	0.722

Best parameters for each model was determined by grid search and all models ended up with the same hyperparameters for the best model.

- C = 100
- penalty = l2
- Solver = Newton-CG

Additional Models - Logistic Regression

Global Coefficients

	Features	Low	Medium	High
52	Israel	-2.386244	-0.330122	2.716366
72	United States	-2.634347	0.099896	2.534451
68	Switzerland	-2.907593	0.465069	2.442524
58	Norway	-4.418748	2.067867	2.350881
42	Denmark	-1.591964	0.212040	1.379924
51	Ireland	-3.672717	2.309114	1.363603
35	Belgium	-2.526257	1.360440	1.165817
33	Australia	-1.751040	0.586666	1.164373
57	New Zealand	-1.971421	1.080935	0.890486
74	Yearly	-1.049810	0.220143	0.829667

Conclusion

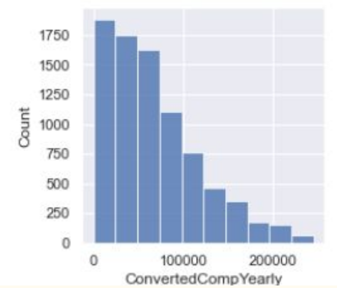
Low correlation in features with outcome variables lead to lower predictive power

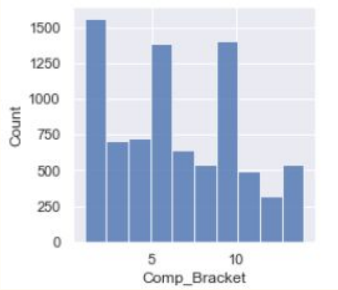
Best Practical Model:
Linear Regression with XGBoost

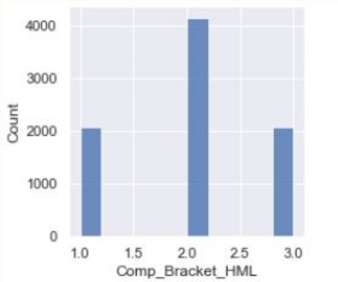
R^2: 0.656
RMSE: $29,166.97

Using the generated compensation brackets did not improve model predictions

High F1 scores with high, medium, low outcome variables







Limitations and Future Work

- Look to improve RMSE to be less than $29,166.97
- We would like to collect more data
  - Industry
  - State/Metropolitan Area
  - Hours worked
- Narrow scope to only full-time
- Potentially try to develop different outcome brackets

Q&A

Appendix

Appendix - Contribution

- Mai La:
  - Data cleaning & processing: 2.2. Skills, 2.3. Countries & compensation frequency
  - EDA: 4.1. Compensation distribution, 4.2. Skills distribution, 4.3. Features distribution
  - Model data: 5.1. All countries. Model Training - Continuous Target : Step 6
  - Report: Initial writing, Project Summary & Conclusion. Presentation: Slides 8-11
- Matt Pribadi:
  - Data cleaning: 2.1. Cleaned up Years Programmed (professionally and amature), Age, organizational size; Developed framework for functions
  - Modeling: 7.1 to 7.2. RandomForestClassification Model, US and Global data, Important Features EDA, Tree printing
  - Presentation: Random Forest Model & Conclusion. Slides 12, 17
  - Report: Editing
- Charlie Boatwright:
  - Data Cleaning: 2.3 Categorical Features Ethnicity, Education, Gender, Sexual Orientation, Employment status
  - Modeling: 7.3 US and Global categorical modeling with Logistic Regression and analysis
  - Presentation: Introduction, EDA, (slides 1, 2, 3, 7) Logistic Regression slides 13 and 14
  - Report: Editing
- Jacquie Nesbitt:
  - Data Master: 3 - 3.2 Made starting master data document combining 3 years of survey data, matched columns
  - Data Cleaning: 2.1, also built the categorical outcome variables for the categorical models
  - Modeling: 7.4 US and Global categorical modeling for SVM Linear and SVM Radial Basis Function
  - Presentation: Build outline for baseline and final presentation, SVM Model and Limitations. Slides 13, 14, 18
  - Report: Edits and responsible for submission
  - Project Management: team notes, meetings, timeline management

Algorithms

Continuous Outcome Variable

- Linear Regression
  - Log Transform Compensation
  - RandomForestRegressor, SVR

```
## Base Model - Linear Regression:
MSE train: 2424428011.456
MSE test: 2383543534.743

R2 Score train: 0.429
R2 Score test: 0.445
```

```
## Linear Regression - Transform Compensation to Log scale Model:
MSE train: 0.380
MSE test: 0.362

R2 Score train: 0.622
R2 Score test: 0.638
```

Categorical Outcome Variable

- Logistic Regression
- Decision Trees/Random Forest
- SVM
- Ensemble

Algorithms

Only look at US data
Re-aggregate education - Re aggregate everything below a college degree (associates) or throw them away or impute with mode
Build another categorical outcome variable (High, medium, and low earners)

Baseline variables to use:

- Num of languages and num of languages for data science (Mai's created a second grouping)
- Codepro
- Age1stcode
- Orgsize
-

Continuous Outcome Variable

- Linear Regression
    - Log Transform Compensation
    - RandomForestRegressor, SVR

Categorical Outcome Variable

- Logistic Regression
- Decision Trees/Random Forest
- SVM
- Ensemble

Evaluation

- Regression:
  - Adjusted R-squared to compare different model options

- Classification:
  - F1 score: compensation bracket outcome has class imbalance