MIE1624 Assignment 1 - Predicting Kaggle Data Scientist Compensations

Alex Kwan - 1001559057

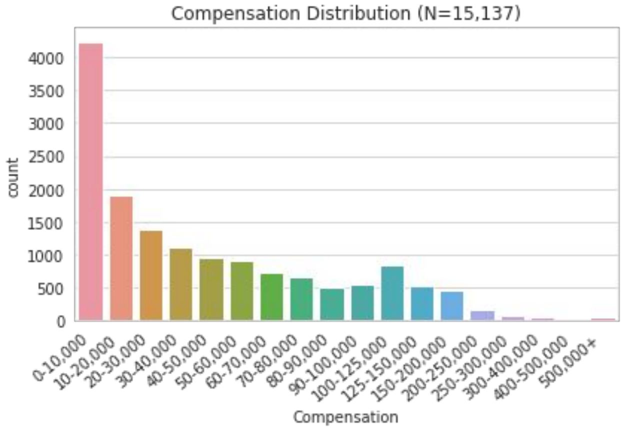
Introduction & Exploratory Data Analysis

Goal: Predict annual compensation range in USD for a given data scientist given various features including: Gender, Age, Country, Education, Industry, Years Experience, etc

→ Ordinal classification problem since target is categorical with 18 ranges from 0-10k to 500k+

Raw Dataset: Kaggle Survey with 15,429 samples and 314 columns

Cleaned Dataset: 15,137 samples with a target label

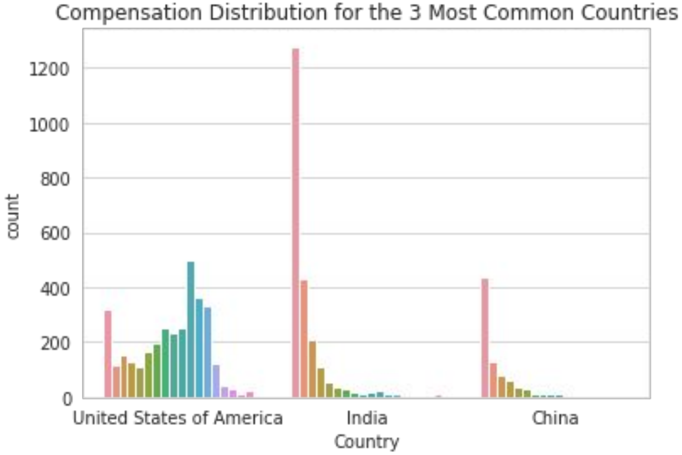


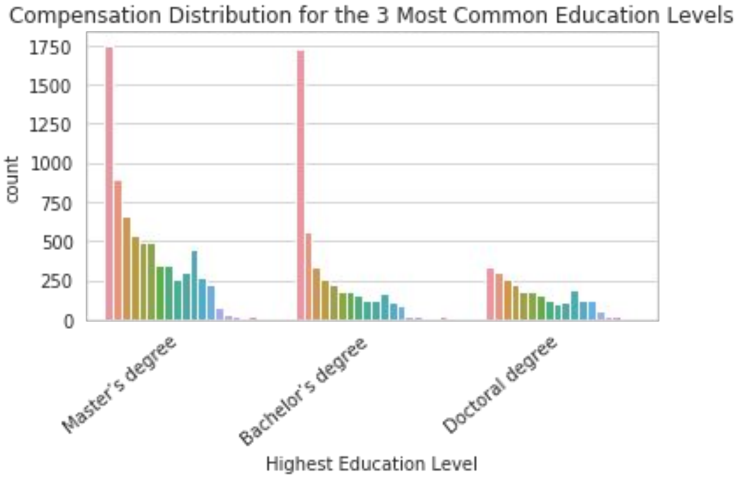
Note: Target variable is skewed to lower compensations

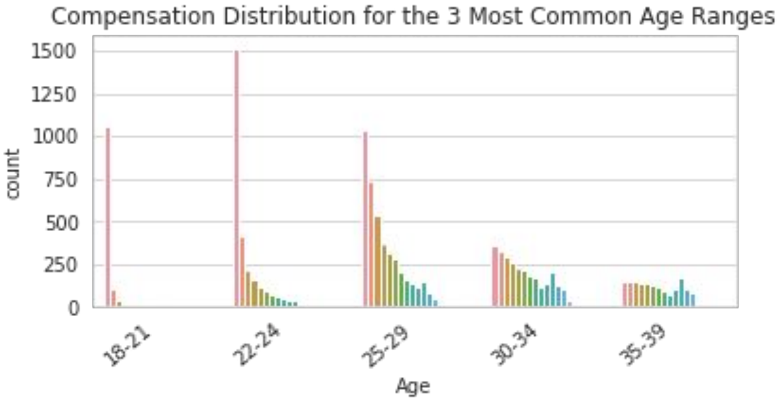
Visualizations

Compensation is higher for those who:

- live in America as opposed to India or China
- have a high level of education
- are older

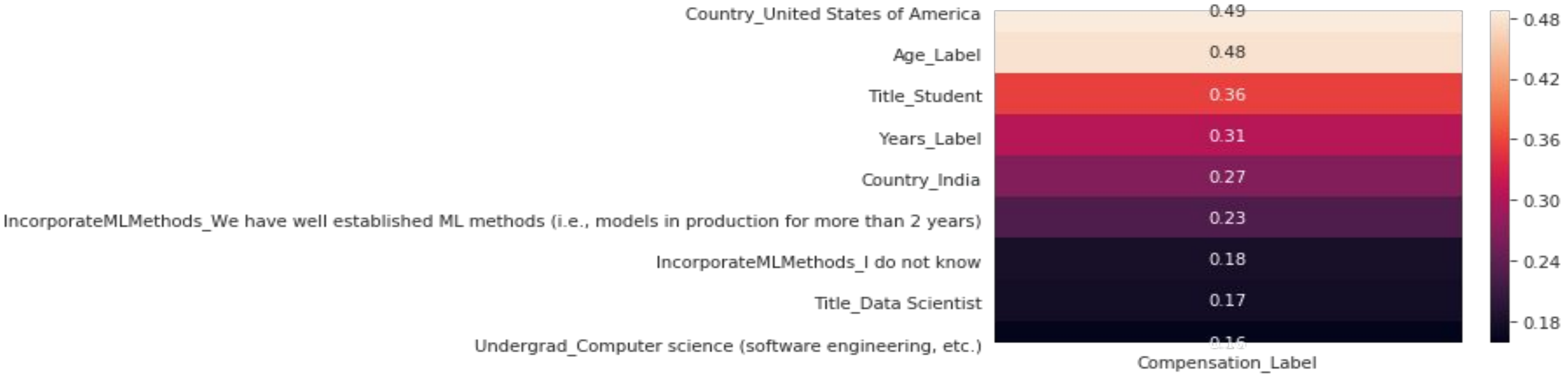






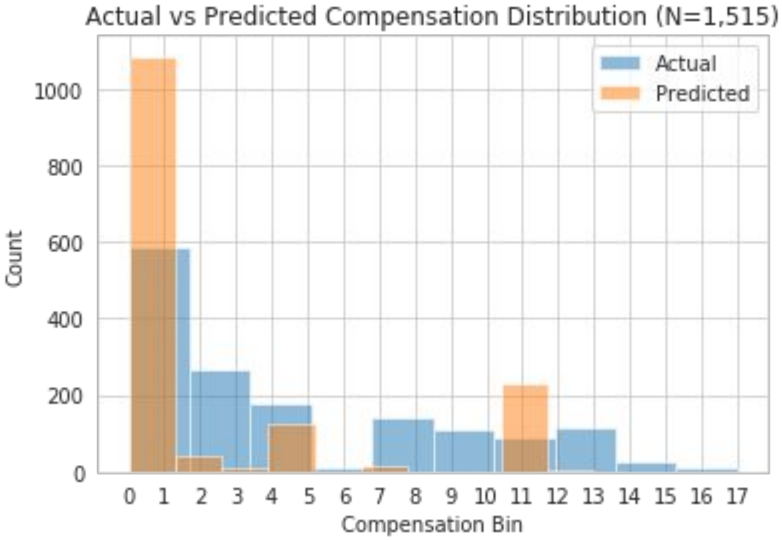
Model Feature Importance

- Filled NaN values with 0 for existing binary features in the original dataset
- One hot encoded every country and industry since Logistic Regression can’t interpret text directly
- Label-encoded compensation as well as Age and Years Experience to capture the ordered relationship
- Top 10 correlated features with target (absolute correlation):

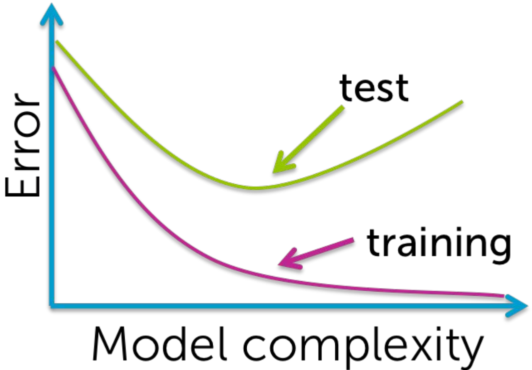


Model Results

- Feature Engineering yields 425 available features
- Feature Selection done automatically with LASSO, C=0.1 → 378 features were kept
- 10 fold CV Grid Search yielded hyperparameters:
    - C = 0.001
    - Solver = newton-cg
- Optimal Logistic Model 10 fold CV performance
    - Average Accuracy (F1-score): 33.442%
    - Average Standard Deviation: 5.7%
- Clearly room for improvement
    - Model tends to over-predict the lowest compensation bin and is unable to accurately predict high compensation bins



Discussion and Future Work



Overfitting or Underfitting?

- Since training set accuracy (34%) and test accuracy (32%) are close, I don't believe the model is overfitting
- Only way to verify underfitting would be to plot the Error/Accuracy vs. Model complexity
  - Low model complexity would cause underfitting → low values of C means the regularization strength is high
  - Error is composed of two components the squared bias and the variance (want to minimize both)
  - Use F1-score to measure accuracy since it considers both precision and recall

Future Work:

- Test more C values in the Grid Search and plot Error/Accuracy vs. Model Complexity
- Use PCA to reduce feature dimensionality or tune C for LASSO feature selection
- Reduce the cardinality of the target variable from 18 compensation bins to >10 bins
- Select a subset of data points (focus on a specific country or age bin)