Algorithmic Bias in AI Resume Screening

Aidan Bugayong

November 14, 2025

Table of contents

1 Introduction

1.1 Context and Background

Resume screeners were developed in order to screen canidates more efficiently and reduce the human bias in the screening process. However, there are concerns with whether or not these automated resume screening systems are truly unbiased in their decision making process. As more and more companies use some form of AI automation in their hiring process, the question of whether or not these systems are unbiased has become more important.

According to Naveen Kumar in his article on AI recruitment statistics, roughly 87% of companies are using AI in the hiring and recruitment process[1]. As such, the University of Washington decided to conduct their own study to determine if there is any bias in the decision making process of resume screeners and what sorts of factors contribute to the scoring process of a resume. Their research found "significant racial, gender and intersectional bias in how three state-of-the-art large language models, or LLMs, ranked resumes. The researchers varied names associated with white and Black men and women across over 550 real-world resumes and found the LLMs favored white-associated names 85% of the time, female-associated names only 11% of the time, and never favored Black male-associated names over white male-associated names"[2]. Their research came to these conclusions by handing the AI a list of identical resumes with different names and then having the AI give them scores. The article ends with the research team noting that more research should be done in this area by looking at different attributes and more LLMs in order to better allign these AI systems with the real world policies to reduce bias and harm.

1.2 Project Purpose

This project aims to determine if there is any bias in the decision making process of resume screeners and what sorts of factors contribute to the scoring process of a resume. More specifically, looking at a wide variety of applicant attributes to determine what factors have the highest contribution to the bias in the decision making process. Ideally, professional experience and education will be the best predictors of the resume scores but we will also look into other factors such as gender, ethnicity, and institutional prestige.

[1] https://www.demandsage.com/ai-recruitment-statistics/

[2] https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/

2 Methods

2.1 Data Collection and Purpose

The original list of resumes is from a dataset in huggingface[3], which is comprised of both real and synthetic resume data in JSON formatting. The purpose of this dataset was for training natural language processing (NLP) models for resume parsing. Specifically, the resumes are oriented around technical roles and is designed for NLP models to be trained on and used for candidate matching / screening in this field. This dataset was posted on huggingface Feburary 21st, 2025 and has not been updated since (excluding the minor changes to the readme).

The sources used for this dataset are sourced from anonymized CV submissions as well as synthetiic resumes generated using "Faker Library" and filled with realistic and role appropriate information. All resumes are anonymized by removing PII (Personally Identifiable Information) but many fields (such as names) contain realistic placeholders. The makers of the dataset note that the data is oriented around technical roles and the synthetic resumes may not capture the same nuances of a real resume. As such, the makers note that this dataset should only be used for NLP, data augmentation, or exploratory data analysis and should not be used for non-technical roles or personalized hiring decisions.

The dataset contains over 4500 resumes in a JSON format. Each resume entry contains personal information, work experience, education information, skills and projects. Since these are technical resumes (oriented around the computer science / information technology field), the skills and projects fields contained a candidate's coding projects and/or coding languages. For the scope of this project, all fields were used for analysis or scoring of the resume.

2.2 Data Processing

The collection of resumes was processed in order to create the score datasets used for this project. The score datasets has the following columns: Name (acting as the primary key), Resume Score, Gender, Ethnicity, Institutional Prestige, Years of Experience, Skill Relevance, Experience Relevance and Project Relevance. The score datasets are specific to each of the langauge models used in this project. This is because we are trying to understand the relation between how the model scores a reseume and the model's own perception of the candidate (gender, ethnicity, and institutional prestige).

[3]https://huggingface.co/datasets/datasetmaster/resumes

For creating the score datasets themselves, the resumes were scored by the model and then the model was asked to guess the gender, ethnicity, and institutional prestige of the candidate. For determining the skill/project/expperience relevance, that was also a call to the language model. It is important to note that each column of data is a separate instance of the language model to reduce the liklihood of previous responses affecting the current responses. All of this information was then used to create the score dataset for each of the models used in this project.

The models used for this project include IBM's *granite3.3:8b*, Microsoft's *Phi4:14b* and Meta's *llama3.1:8b*. These models were choosen because they are highly rated models despite the lower paramter size and are all open source.

2.3 Statistical Tools and Approach

2.3a Exploratory Data Analysis

After the resumes have been processed by the language models, there was some data cleaning to ensure the models followed the instructions. Once the data has been cleaned, exploratory data analysis was conducted to get a better understanding of the data distributions and relationships. Pi charts were used to show the distribution of the categorical variables and histograms were used to show the distribution of the numeric variables for each of the datasets (IBM, Microsoft, and Meta).

2.3b Initial Modeling Approach

The initial approach was to use linear regression to predict the resume score from all of the other variables in the dataset. The formula used for this model was: `score ~ gender + ethnicity + prestige + skill_score + project_score + experience_score + years_experience`. Here, the response variable is the resume score which are integers in the range 0-100. The categorical variables are: gender, ethnicity, and prestige. The numeric variables are: years of experience, skill score, project score, and experience score. All categorical variables were setup such that "Unknown" was the reference category. The numeric variables, like the resume score, are in range 0-100.

2.3c Model Diagnostics

This is where there was analysis of interation effects and determining if any interactions were considered statistically significant. If the interaction was considered significant, then the interaction was added to the model. Additionally, the model diagnostics were conducted to ensure that the model's assumptions were met. Specifically, the model was checked for normality, homoscedasticity, and multicollinearity by using the Shapiro-Wilk test, Browne-Forsythe test, and variance inflation factor, respectively.

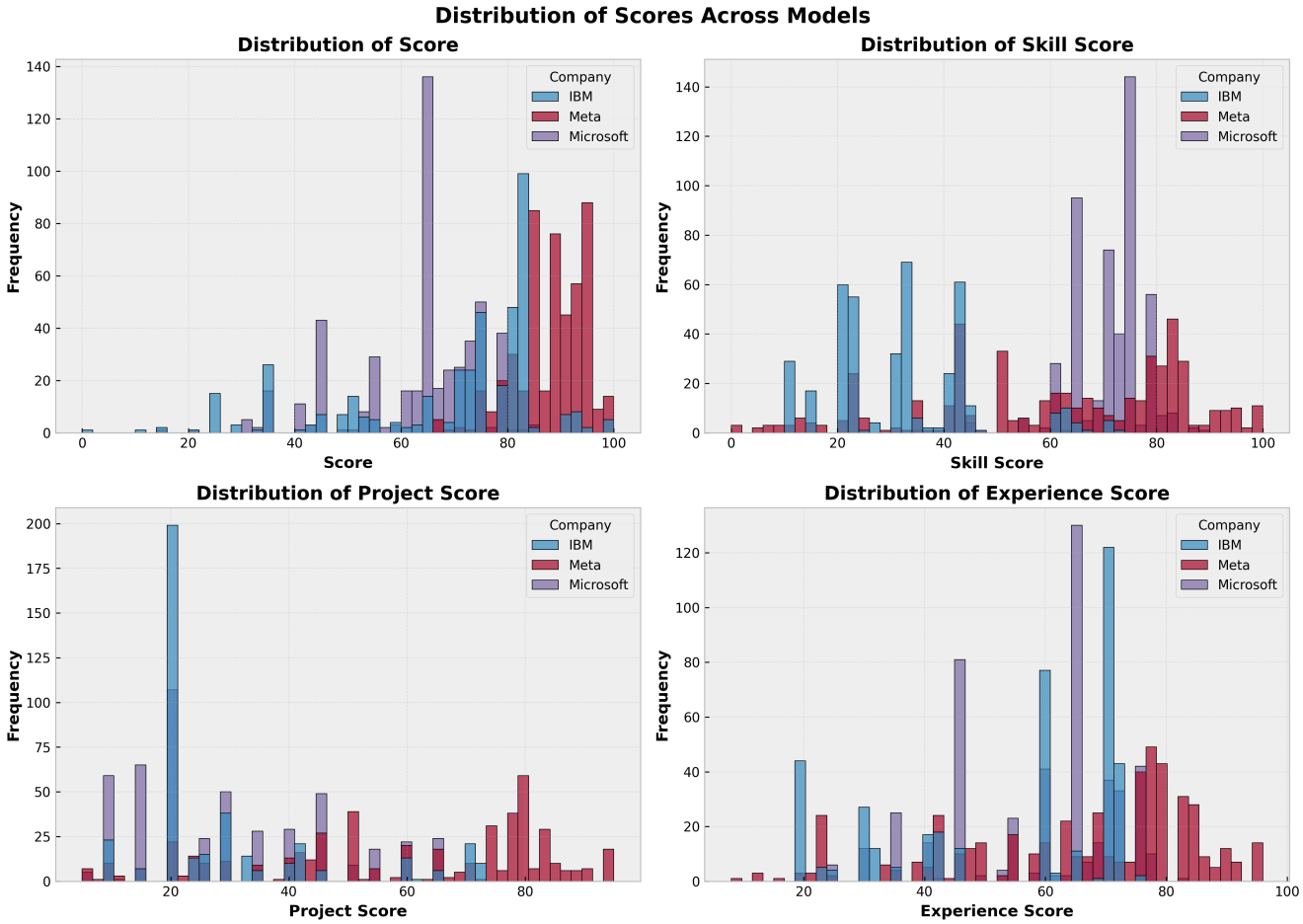
2.3d Model Refinement

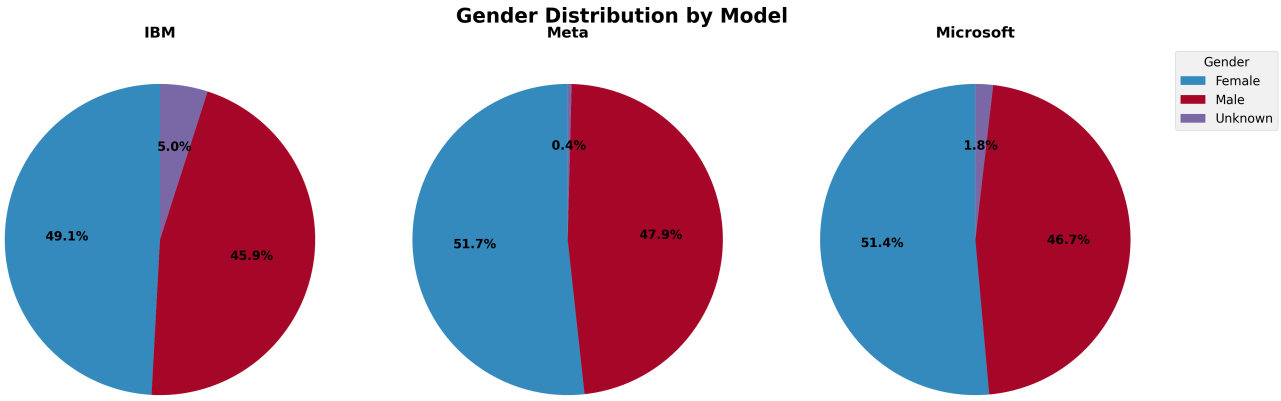
Here, interaction effects were added to the model if they were statistically significant. Due to multicolinearity issues, the model used a ridge regression.

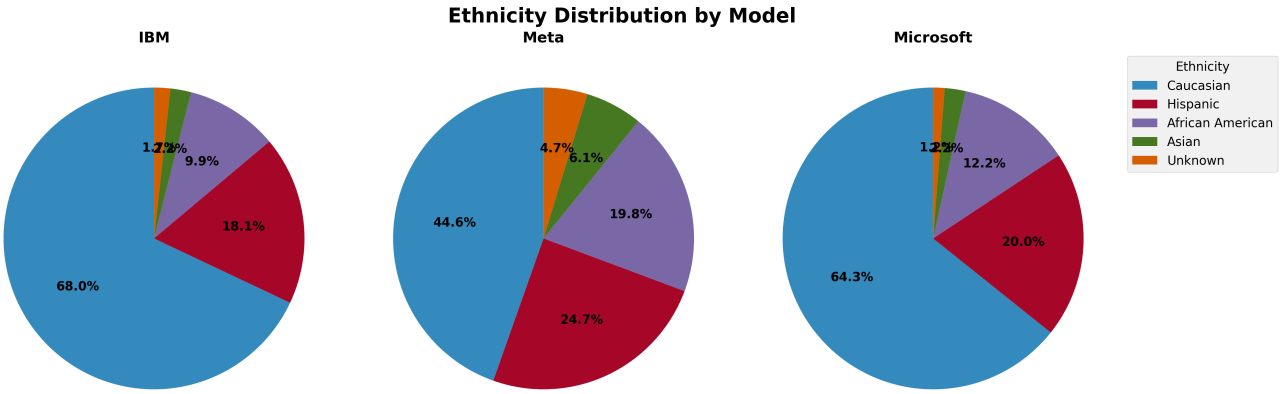
3 Results and Discussion

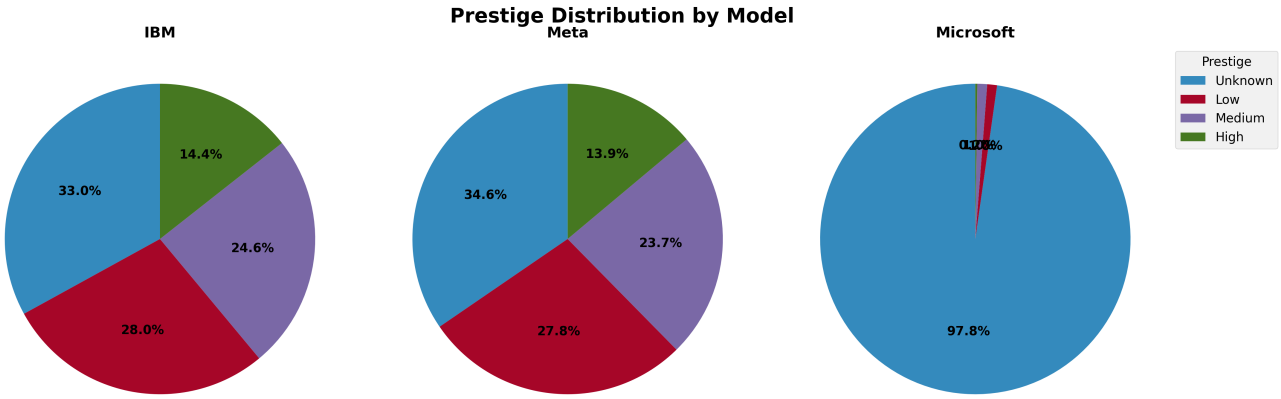
Note that the results are not final and are subject to change. Additionally, the format will be updated

3.1 Data Exploration









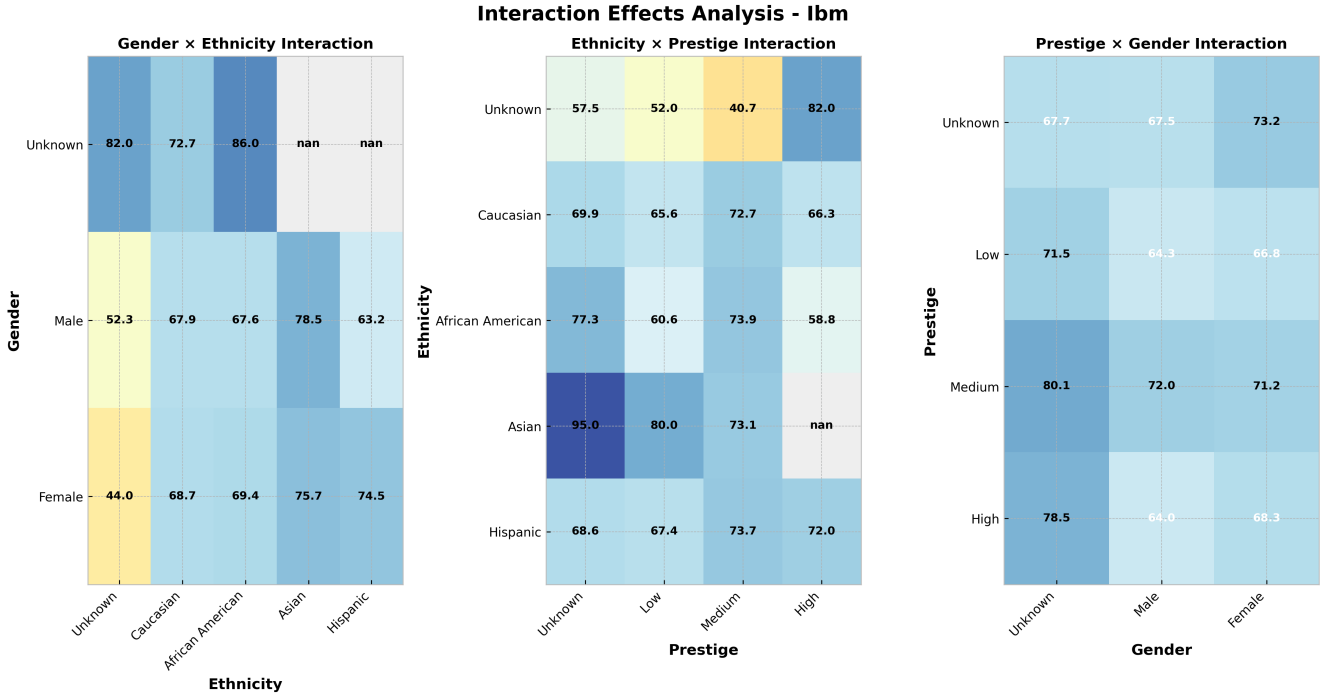
3.2 Initial Model Development

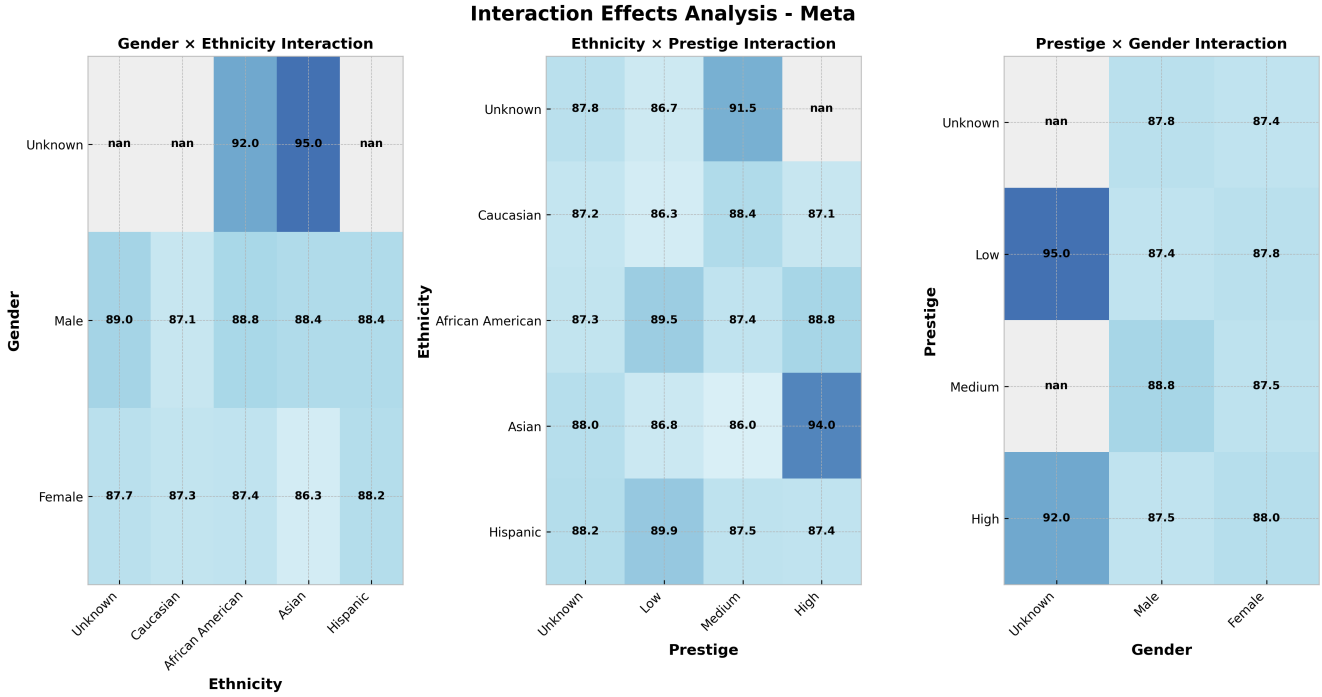
```
Created initial model for ibm
 R-squared: 0.0920
Created initial model for meta
```

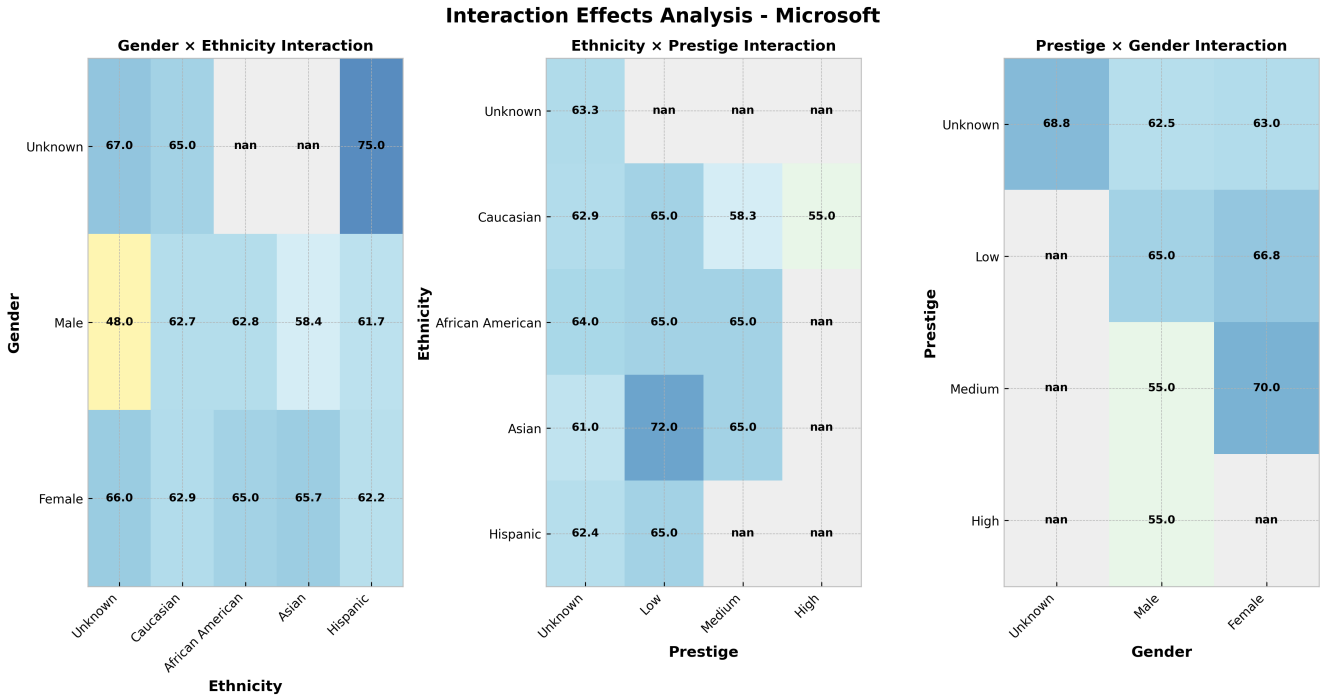
```
  R-squared: 0.0170
Created initial model for microsoft
  R-squared: 0.3371
```

3.3 Interaction Effects and Model Refinement

```
Testing interactions for ibm:
  gender × ethnicity: p = 0.2095 (not significant)
  ethnicity × prestige: p = 0.3570 (not significant)
  prestige × gender: p = 0.6700 (not significant)

Testing interactions for meta:
  gender × ethnicity: p = 0.8907 (not significant)
  ethnicity × prestige: p = 0.3740 (not significant)
  prestige × gender: p = 0.7642 (not significant)

Testing interactions for microsoft:
  gender × ethnicity: p = 0.4816 (not significant)
  ethnicity × prestige: p = 0.9220 (not significant)
  prestige × gender: p = 0.8920 (not significant)
```



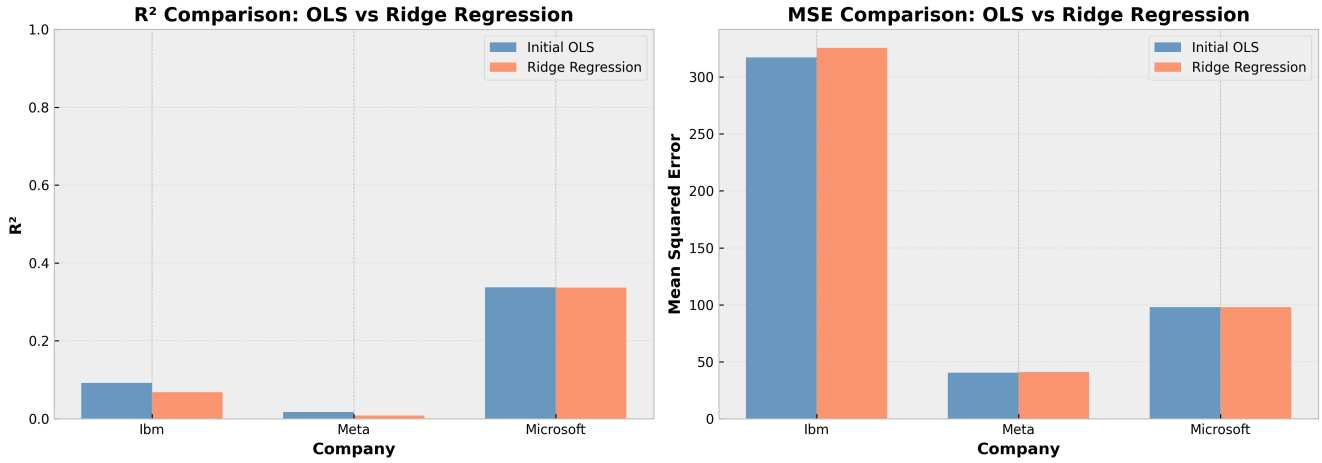





```
Checking assumptions for ibm model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.0006 (Not Heteroscedastic)
  Multicollinearity: 2 variables with VIF > 10
```

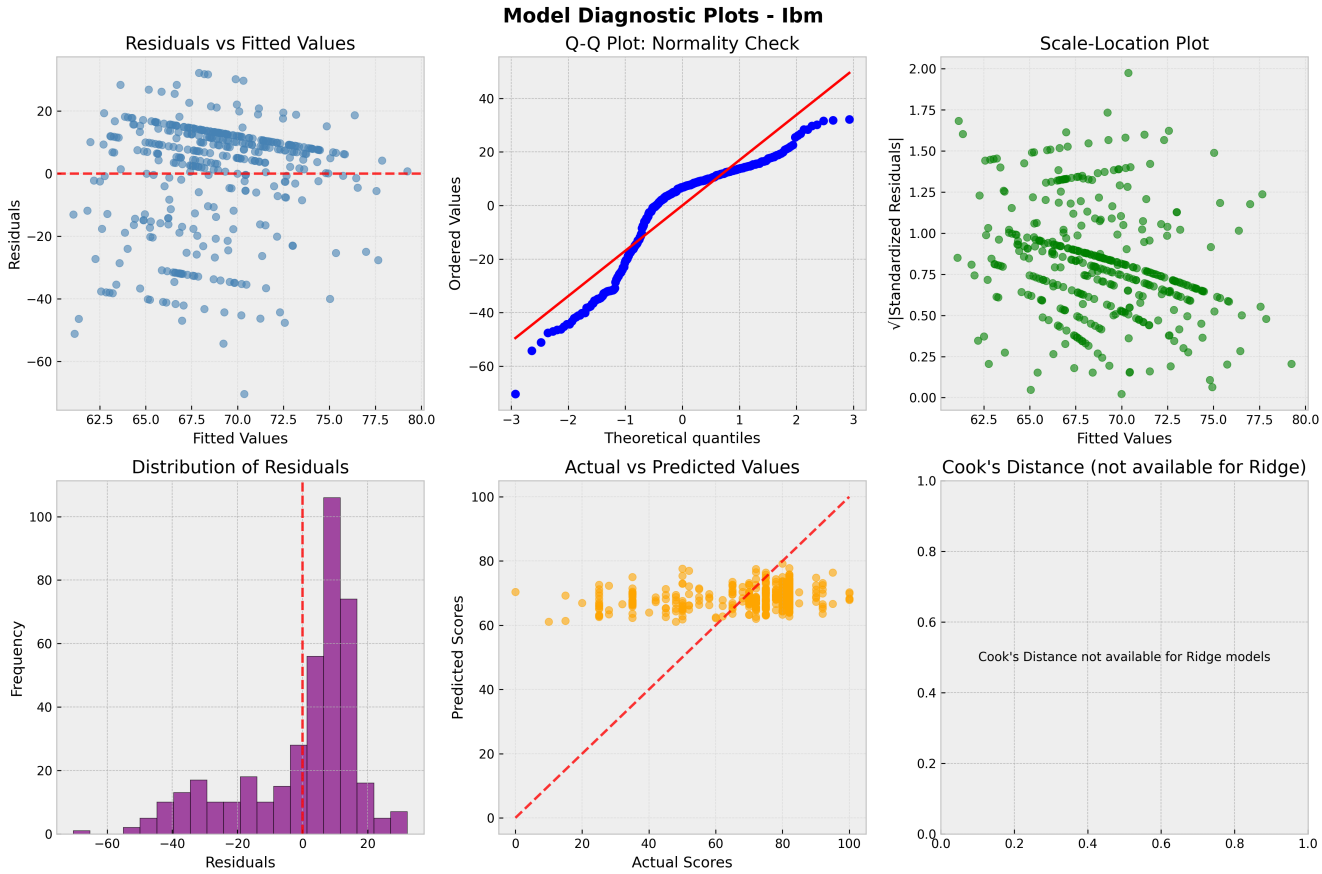
```
  Outliers: 18 potential outliers detected

Checking assumptions for meta model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.1408 (Homoscedastic)
  Multicollinearity: 3 variables with VIF > 10
  Outliers: 9 potential outliers detected

Checking assumptions for microsoft model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.0000 (Not Heteroscedastic)
  Multicollinearity: 6 variables with VIF > 10
  Outliers: 16 potential outliers detected


Fitting Ridge Regression for ibm:
  Optimal alpha: 278.2559
  Initial OLS R²: 0.0920
  Ridge R²: 0.0678
  R² Improvement: -0.0243
  Initial MSE: 317.1882
  Ridge MSE: 325.6599
  MSE Improvement: -8.4717
  Number of features: 13

Fitting Ridge Regression for meta:
  Optimal alpha: 1000.0000
  Initial OLS R²: 0.0170
  Ridge R²: 0.0078
  R² Improvement: -0.0092
  Initial MSE: 40.4270
  Ridge MSE: 40.8042
  MSE Improvement: -0.3772
  Number of features: 13

Fitting Ridge Regression for microsoft:
  Optimal alpha: 10.7227
  Initial OLS R²: 0.3371
  Ridge R²: 0.3363
  R² Improvement: -0.0009
  Initial MSE: 97.8146
  Ridge MSE: 97.9435
  MSE Improvement: -0.1289
```

```
  Number of features: 13
```

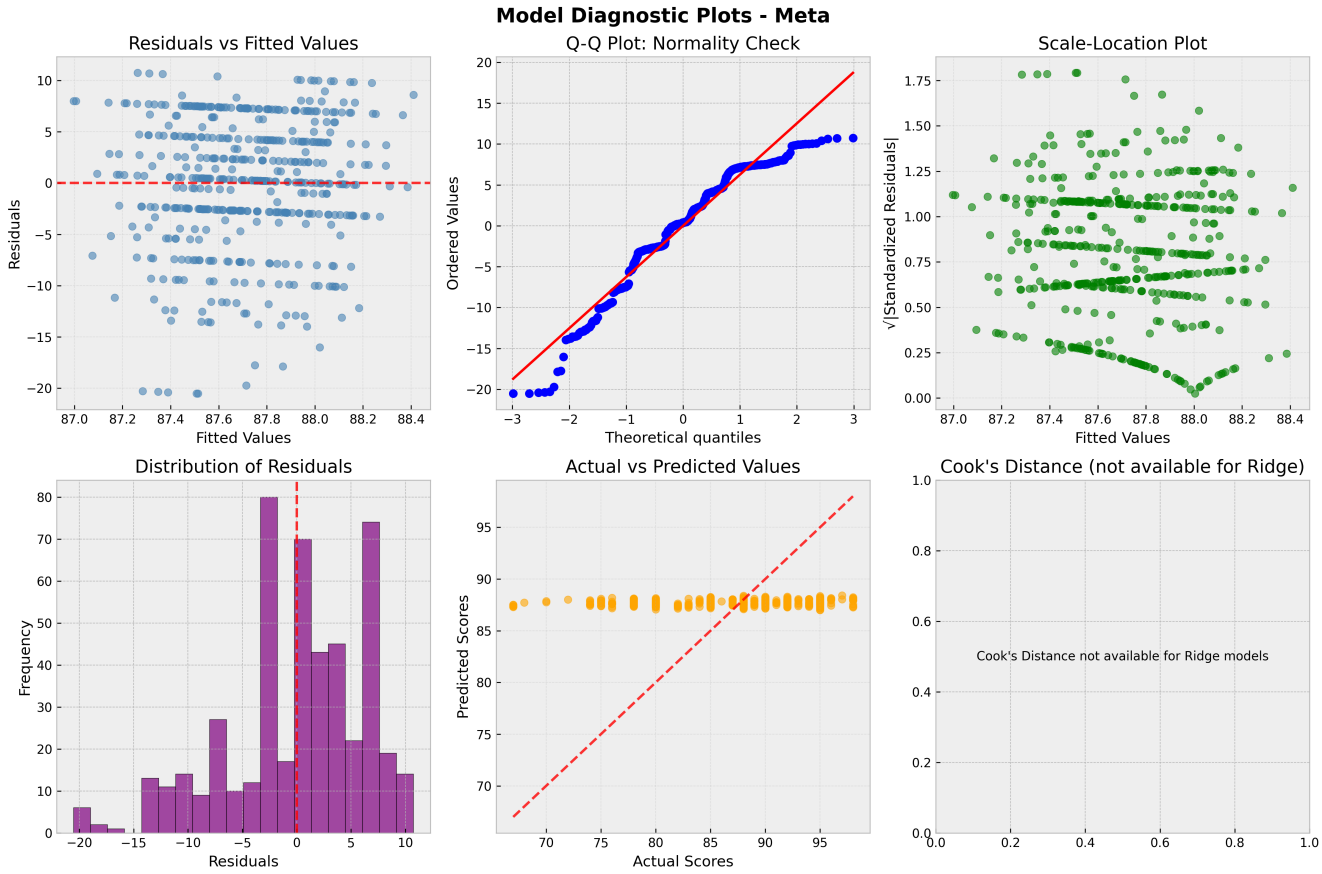


3.4. Model Diagnostics and Assumptions

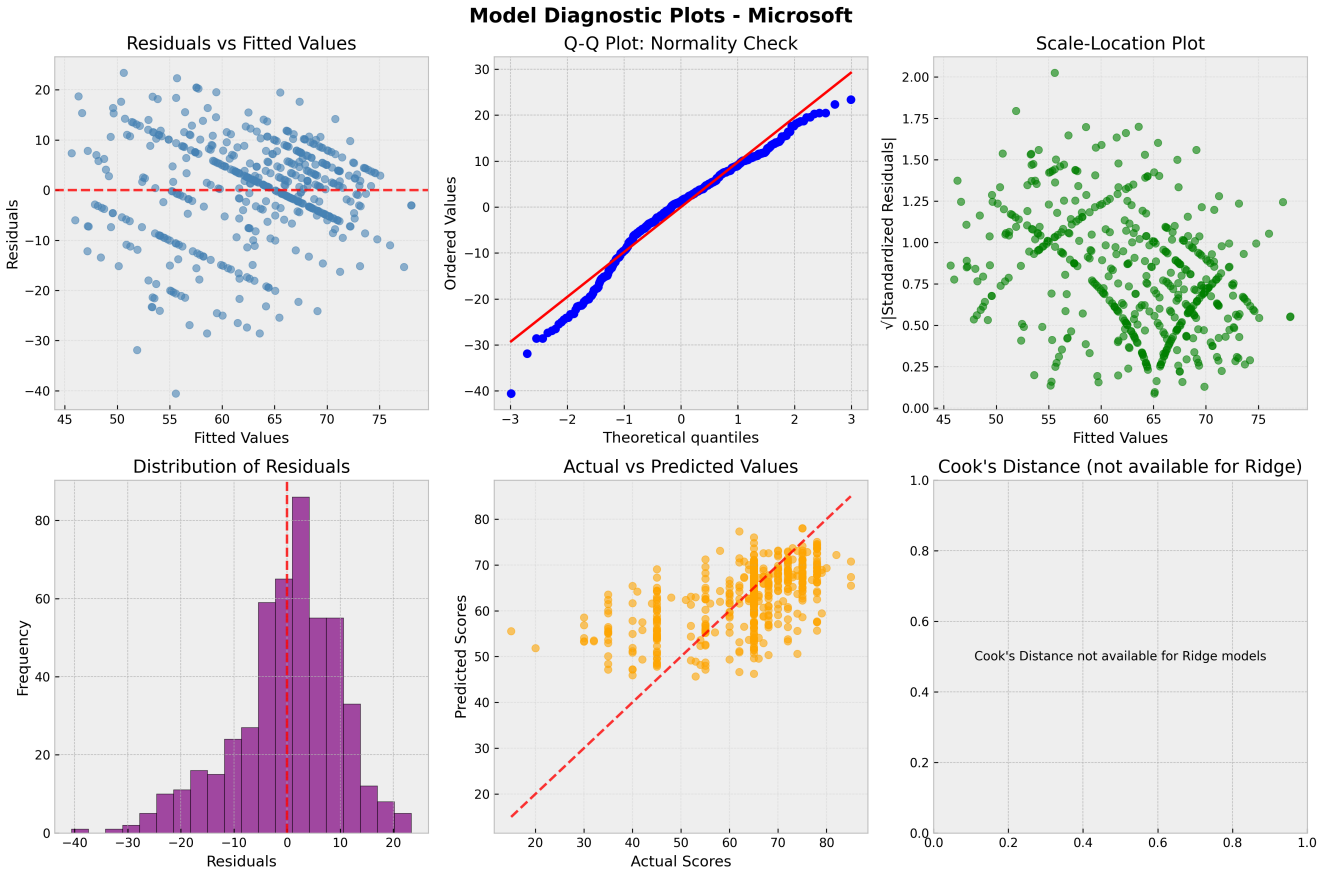
```
Checking assumptions for ibm model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.0008 (Not Heteroscedastic)
  VIF calculation failed: ufunc 'isfinite' not supported for the input types, and the inpu
  Outliers: 32 potential outliers detected

Checking assumptions for meta model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.1236 (Homoscedastic)
  VIF calculation failed: ufunc 'isfinite' not supported for the input types, and the inpu
  Outliers: 9 potential outliers detected

Checking assumptions for microsoft model:
  Normality (Shapiro-Wilk): p = 0.0000 (Non-normal)
  Homoscedasticity (Brown-Forsythe): p = 0.0000 (Not Heteroscedastic)
  VIF calculation failed: ufunc 'isfinite' not supported for the input types, and the inpu
  Outliers: 16 potential outliers detected


Creating diagnostic plots for ibm...
```



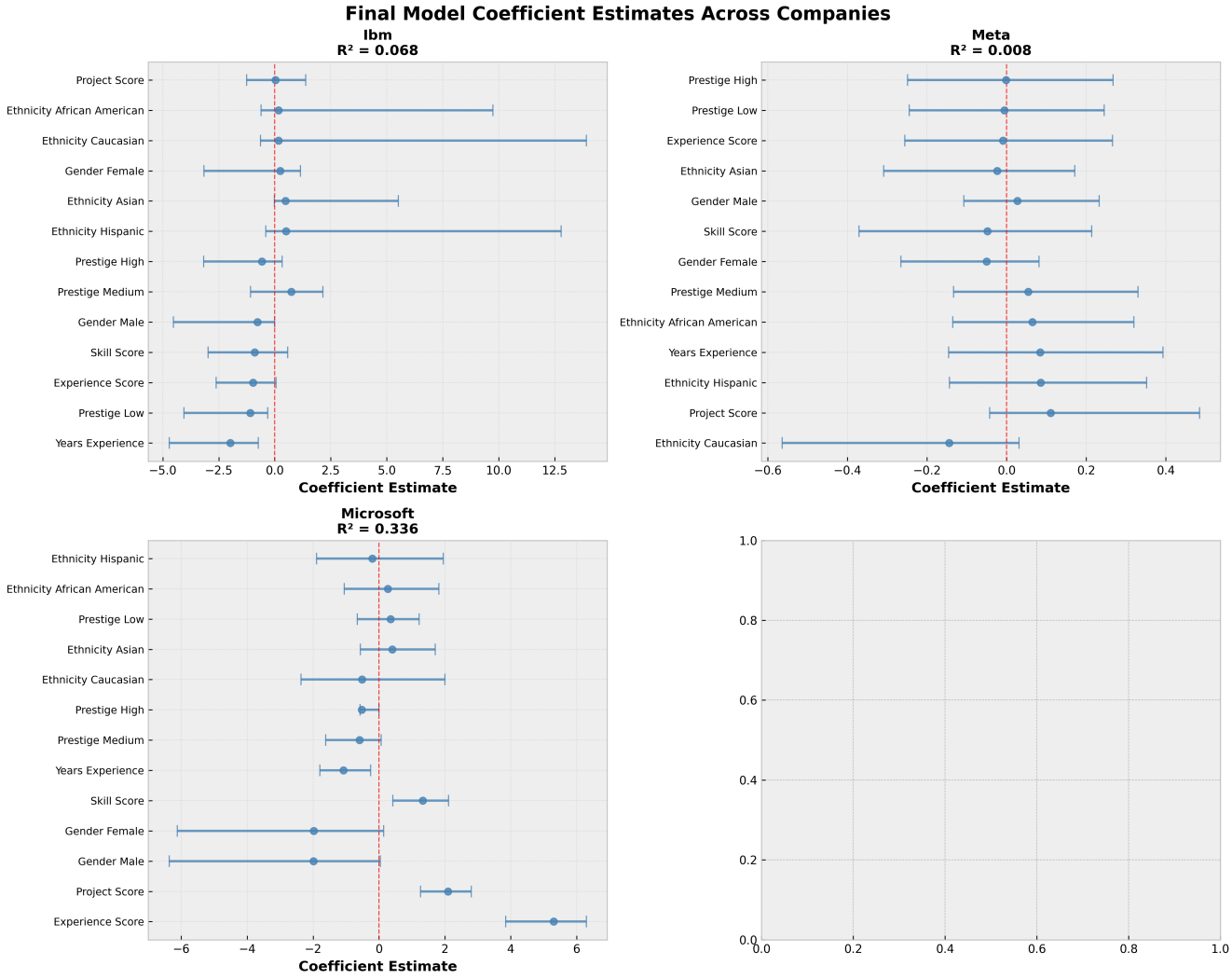

```
Creating diagnostic plots for meta...
```




```
Creating diagnostic plots for microsoft...
```


Model Diagnostic Plots - Microsoft
Residuals vs Fitted Values
Residuals
Fitted Values
Q-Q Plot: Normality Check
Ordered Values
Theoretical quantiles
Scale-Location Plot
√|Standardized Residuals|
Fitted Values
Distribution of Residuals
Frequency
Residuals
Actual vs Predicted Values
Predicted Scores
Actual Scores
Cook's Distance (not available for Ridge)
Cook's Distance not available for Ridge models

3.5 Final Model Interpretation and Comparisons

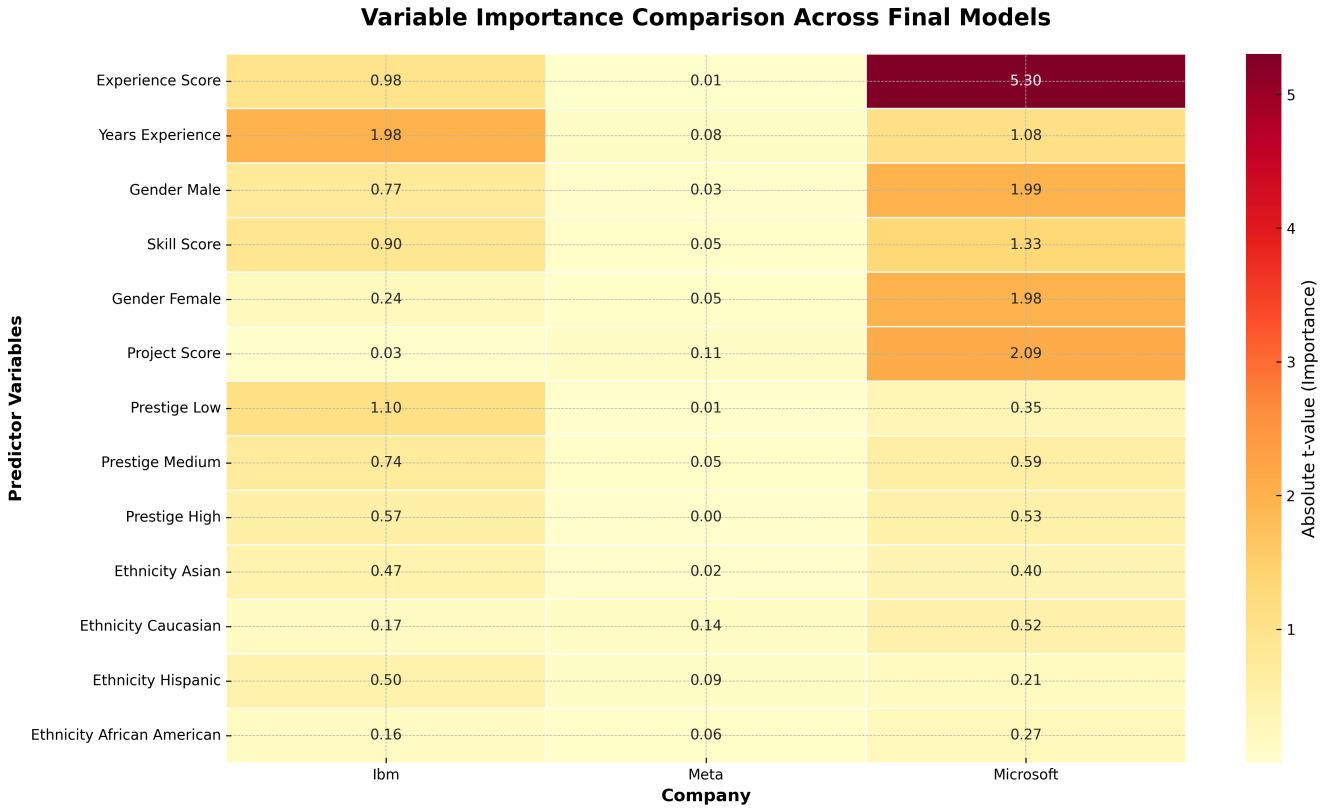



```
FINAL MODEL SUMMARIES:
------------------------------------------------------------

IBM Final Model:
  Model Type: Ridge Regression
  Optimal Alpha: 278.2559
  R-squared: 0.0678
  Number of observations: 403
  Number of features: 13
  Initial OLS Coefficients and p-values:
    Intercept: +70.8426 (p = 0.0000)
```

```
    C(gender, Treatment(reference="Unknown"))[T.Male]: -5.3465 (p = 0.2233)
    C(gender, Treatment(reference="Unknown"))[T.Female]: -2.7226 (p = 0.5342)
    C(ethnicity, Treatment(reference="Unknown"))[T.Caucasian]: +16.5297 (p = 0.0185)
    C(ethnicity, Treatment(reference="Unknown"))[T.African American]: +16.8855 (p = 0.0246
    C(ethnicity, Treatment(reference="Unknown"))[T.Asian]: +20.6559 (p = 0.0261)
    C(ethnicity, Treatment(reference="Unknown"))[T.Hispanic]: +18.6021 (p = 0.0106)
    C(prestige, Treatment(reference="Unknown"))[T.Low]: -4.5055 (p = 0.0645)
    C(prestige, Treatment(reference="Unknown"))[T.Medium]: +1.2315 (p = 0.6264)
    C(prestige, Treatment(reference="Unknown"))[T.High]: -3.7050 (p = 0.2208)
    skill_score: -0.0967 (p = 0.1607)
    project_score: +0.0027 (p = 0.9627)
    experience_score: -0.0698 (p = 0.1736)
    years_experience: -0.8837 (p = 0.0006)
  Ridge Coefficients (approx. p-values from bootstrap and 95% CI):
    skill_score: -0.8995 (p  0.0800) [-2.0150, 0.0563]
    project_score: +0.0283 (p  0.9933) [-0.9608, 0.8812]
    experience_score: -0.9774 (p  0.0400) [-1.8883, -0.0532]
    years_experience: -1.9836 (p  0.0000) [-2.9930, -0.8741]
    gender_Male: -0.7686 (p  0.0867) [-1.4909, 0.1124]
    gender_Female: +0.2430 (p  0.5333) [-0.4568, 0.9568]
    ethnicity_Caucasian: +0.1724 (p  0.7733) [-0.6602, 0.9852]
    ethnicity_African American: +0.1647 (p  0.6800) [-0.6568, 1.0268]
    ethnicity_Asian: +0.4737 (p  0.1400) [-0.1549, 1.0644]
    ethnicity_Hispanic: +0.4971 (p  0.2467) [-0.4692, 1.3205]
    prestige_Low: -1.0970 (p  0.0467) [-2.1098, -0.0435]
    prestige_Medium: +0.7410 (p  0.1200) [-0.1571, 1.4808]
    prestige_High: -0.5693 (p  0.2600) [-1.6966, 0.4145]

META Final Model:
  Model Type: Ridge Regression
  Optimal Alpha: 1000.0000
  R-squared: 0.0078
  Number of observations: 489
  Number of features: 13
  Initial OLS Coefficients and p-values:
    Intercept: +93.6229 (p = 0.0000)
    C(gender, Treatment(reference="Unknown"))[T.Male]: -5.8741 (p = 0.2095)
    C(gender, Treatment(reference="Unknown"))[T.Female]: -6.3238 (p = 0.1771)
    C(ethnicity, Treatment(reference="Unknown"))[T.Caucasian]: -1.1923 (p = 0.3490)
    C(ethnicity, Treatment(reference="Unknown"))[T.African American]: -0.1538 (p = 0.9102)
    C(ethnicity, Treatment(reference="Unknown"))[T.Asian]: -1.0204 (p = 0.5735)
    C(ethnicity, Treatment(reference="Unknown"))[T.Hispanic]: +0.0354 (p = 0.9787)
    C(prestige, Treatment(reference="Unknown"))[T.Low]: +0.0911 (p = 0.9039)
```

```
    C(prestige, Treatment(reference="Unknown"))[T.Medium]: +0.4886 (p = 0.5334)
    C(prestige, Treatment(reference="Unknown"))[T.High]: +0.0254 (p = 0.9785)
    skill_score: -0.0080 (p = 0.5175)
    project_score: +0.0146 (p = 0.2291)
    experience_score: -0.0028 (p = 0.8499)
    years_experience: +0.0602 (p = 0.4542)
  Ridge Coefficients (approx. p-values from bootstrap and 95% CI):
    skill_score: -0.0480 (p  0.5533) [-0.2358, 0.1231]
    project_score: +0.1106 (p  0.2133) [-0.0615, 0.2872]
    experience_score: -0.0089 (p  0.8800) [-0.2047, 0.1952]
    years_experience: +0.0844 (p  0.4933) [-0.0841, 0.2664]
    gender_Male: +0.0272 (p  0.6533) [-0.1183, 0.1712]
    gender_Female: -0.0502 (p  0.4600) [-0.1859, 0.1014]
    ethnicity_Caucasian: -0.1441 (p  0.1000) [-0.3006, 0.0145]
    ethnicity_African American: +0.0647 (p  0.4667) [-0.0836, 0.2188]
    ethnicity_Asian: -0.0234 (p  0.8867) [-0.2106, 0.1389]
    ethnicity_Hispanic: +0.0859 (p  0.3067) [-0.0860, 0.2351]
    prestige_Low: -0.0056 (p  0.8800) [-0.1630, 0.1751]
    prestige_Medium: +0.0543 (p  0.4867) [-0.1274, 0.2343]
    prestige_High: -0.0016 (p  0.9600) [-0.1884, 0.1852]

MICROSOFT Final Model:
  Model Type: Ridge Regression
  Optimal Alpha: 10.7227
  R-squared: 0.3363
  Number of observations: 490
  Number of features: 13
  Initial OLS Coefficients and p-values:
    Intercept: +29.1747 (p = 0.0001)
    C(gender, Treatment(reference="Unknown"))[T.Male]: -6.5722 (p = 0.0666)
    C(gender, Treatment(reference="Unknown"))[T.Female]: -6.5414 (p = 0.0662)
    C(ethnicity, Treatment(reference="Unknown"))[T.Caucasian]: -0.6342 (p = 0.8829)
    C(ethnicity, Treatment(reference="Unknown"))[T.African American]: +1.3476 (p = 0.7632)
    C(ethnicity, Treatment(reference="Unknown"))[T.Asian]: +3.3896 (p = 0.5228)
    C(ethnicity, Treatment(reference="Unknown"))[T.Hispanic]: -0.1111 (p = 0.9797)
    C(prestige, Treatment(reference="Unknown"))[T.Low]: +3.5797 (p = 0.4329)
    C(prestige, Treatment(reference="Unknown"))[T.Medium]: -6.0239 (p = 0.1879)
    C(prestige, Treatment(reference="Unknown"))[T.High]: -11.9800 (p = 0.2348)
    skill_score: +0.2032 (p = 0.0049)
    project_score: +0.1251 (p = 0.0002)
    experience_score: +0.4155 (p = 0.0000)
    years_experience: -0.3013 (p = 0.0155)
  Ridge Coefficients (approx. p-values from bootstrap and 95% CI):
```

```
skill_score: +1.3279 (p   0.0067) [0.3716, 2.1505]
project_score: +2.0941 (p   0.0000) [1.1366, 2.9665]
experience_score: +5.3002 (p   0.0000) [4.1690, 6.6690]
years_experience: -1.0755 (p   0.0133) [-1.9929, -0.2444]
gender_Male: -1.9864 (p   0.1067) [-3.9808, 0.3122]
gender_Female: -1.9776 (p   0.0933) [-3.8942, 0.1566]
ethnicity_Caucasian: -0.5158 (p   0.6133) [-2.2744, 1.2905]
ethnicity_African American: +0.2700 (p   0.6800) [-1.0183, 1.6399]
ethnicity_Asian: +0.4013 (p   0.5000) [-0.6634, 1.5004]
ethnicity_Hispanic: -0.2055 (p   0.8200) [-1.5427, 1.4471]
prestige_Low: +0.3525 (p   0.4267) [-0.6418, 1.1802]
prestige_Medium: -0.5874 (p   0.1800) [-1.4265, 0.0765]
prestige_High: -0.5268 (p   0.3000) [-0.5883, 0.0000]
```




```
ANALYSIS COMPLETE - SUMMARY

Datasets analyzed: ['ibm', 'meta', 'microsoft']
Total models created: 3

Average R-squared across models: 0.1373
```

```
Most consistently important predictors:
  Experience Score: 2.10
  Years Experience: 1.05
  Gender Male: 0.93
  Skill Score: 0.76
  Gender Female: 0.76

  All analysis complete! Check generated plots for visualizations.
```

4 Conclusion

5 References