Job Satisfaction & Retention Prediction

Business Problem

Technology companies continuously experience challenges retaining developer talent which leads to increased costs, lost knowledge, and disrupted workflows and culture. At the same time, developers want roles that align with their career values, work preferences, and compensation expectations. This analysis uses the Stack Overflow Developer Survey data from 2024 to identify the key predictors of job satisfaction, and the likelihood of job change among developers. These insights can help employers implement strategies to retain talent and assist developers assessing job fit.

Background

Historically, retention efforts have relied on anecdotal evidence or human resources surveys. The Stack Overflow Developer Survey offers a large, self-reported dataset that covers developer experiences globally. Factors such as compensation, remote work preferences, technology stack alignment, and demographic compatibility significantly influence job satisfaction. This project uses modeling techniques to explain how job satisfaction and intent to change jobs can be predicted using the survey data.

Data Explanation

The dataset is sourced from the publicly available 2024 Stack Overflow Developer Survey. It includes 65,437 responses and 114 variables covering demographics, compensation, experience, technology preferences, education, and job satisfaction. The key fields analyzed include JobSat, Employment, ConvertedCompYearly, DevType, YearsCode, RemoteWork, Country, EdLevel, and Age. A classification target variable, JobChangeIntent, was derived from JobSat by assigning scores of five or below as an indicator of intent to leave. Data cleaning involved removing columns with more than fifty percent missing data and filtering out outlier compensation values using the interquartile range method. Categorical variables were encoded through either one-hot or label encoding, and new features were engineered, including experience levels, technology diversity scores, compensation categories, and age groups.

Methods

The analysis followed a structured process consisting of exploratory data analysis, data cleaning, feature engineering, and model training. Two predictive models were developed. The first involved regression modeling of job satisfaction using Linear Regression and Random Forest Regressor. The second used classification modeling to predict intent to leave using Logistic Regression and Random Forest Classifier. Regression models were evaluated using RMSE, MAE, and R-squared, while classification models were assessed through accuracy, F1-score, and ROC-AUC. The data was split using a seventy-fifteen-fifteen distribution for training, validation, and testing with stratified sampling to address a class imbalance.

Analysis

Regression analysis revealed poor performance from the Linear Regression model, which had a RMSE of 2.088 and R-squared of 0.027. The Random Forest Regressor seemed to have an overfitting problem. It benched a training R-squared of 0.859 but a negative validation R-squared of -0.082. The classification models also underperformed. Logistic Regression had a ROC-AUC of 0.644 but failed to predict the minority class with a F1-score of 0.000. Random Forest yielded a ROC-AUC of 0.681 but an F1-score of 0.000 on validation (see Figure 4 in Appendix). Exploratory analysis indicated that sixty percent of developers rated their satisfaction at seven or higher (see Figure 1 in Appendix). Median salary was $65,000, though values extended to as high as $16 million (see Figure 2 in Appendix). The majority demographic was developers aged twenty-five to thirty-four, working remotely or in hybrid roles, mostly in the United States (see Figure 3 in Appendix).

Conclusion

Job satisfaction among developers can be predicted in general terms, however modeling an individual's intent to leave a job is difficult due to class imbalance. Linear models were simply unable to predict, while tree-based models failed to generalize without overfitting. Though, feature engineering showed some indicators, such as experience level, technology stack diversity, and work modality that could help organizations improve retention strategies (see Figure 5 in Appendix).

Assumptions

This analysis assumes that the satisfaction scores provided in the survey are accurate self-reports. It also assumes that compensation values are reported in U.S. dollars and are correct. Developers with satisfaction scores of five or lower were categorized as likely to leave.

Limitations

Only 29,126 of the 65,437 total respondents provided valid satisfaction ratings, limiting the usable dataset for regression. About sixty-six percent of compensation data was missing. The classification target variable was derived based on thresholding rather than from explicitly stated job change intentions. High cardinality in certain categorical features complicated model construction. External validation was not performed beyond the provided dataset split testing. The analysis is restricted to a single year's dataset. Before the final milestone, the analysis will include all available data from past years to account for variable job market conditions.

Challenges

The primary challenge in this analysis was the class imbalance. Only 8.6% of respondents were labeled as likely to leave. This imbalance caused classification models to overwhelmingly favor the majority class. Encoding categorical variables increased feature dimensionality and complexity. The compensation data was highly skewed, which complicated the analysis. The Random Forest models showed significant overfitting, likely due to complex feature interactions.

Future Uses / Additional Applications

This modeling approach could be extended to predict not just binary attrition intent but also degrees of interest in job change, such as passive versus active searching for a new role. Classification performance may be improved using techniques such as SMOTE. A SHAP analysis might help narrow down a better path forward for interpretability. The methods used for this analysis could be used on internal company survey data for specific companies to get more tailored results.

Recommendations

Organizations should monitor metrics like technology diversity and experience level when assessing developer engagement. Surveys that modeled on the JobSat scale should be given every quarter. Predictive models should be continuously updated to account for variable job markets and outputs should be used along with human judgment.

Implementation Plan

Implementation would be developing internal dashboards that display satisfaction predictors and risk scores for specific groups to see if things are going awry with certain teams. Models need to be adjusted for class imbalance and tuned using validation feedback.

Ethical Assessment

There are several ethical concerns in predictive modeling of job satisfaction and retention. Since the survey is self-selected, the model may not generalize to underrepresented groups due to survey sampling bias. Privacy is a major concern and as such, outputs should be anonymized and aggregated to groups and not individuals. Model predictions should not be used to penalize employees or inform hiring decisions without additional context. Transparency is essential for things like evaluating people’s behavior. The models should be used along with interpretability tools and human oversight. The predictions provided by the models should inform but not replace human judgment.

Appendix

Figure 1. Job Satisfaction Distribution

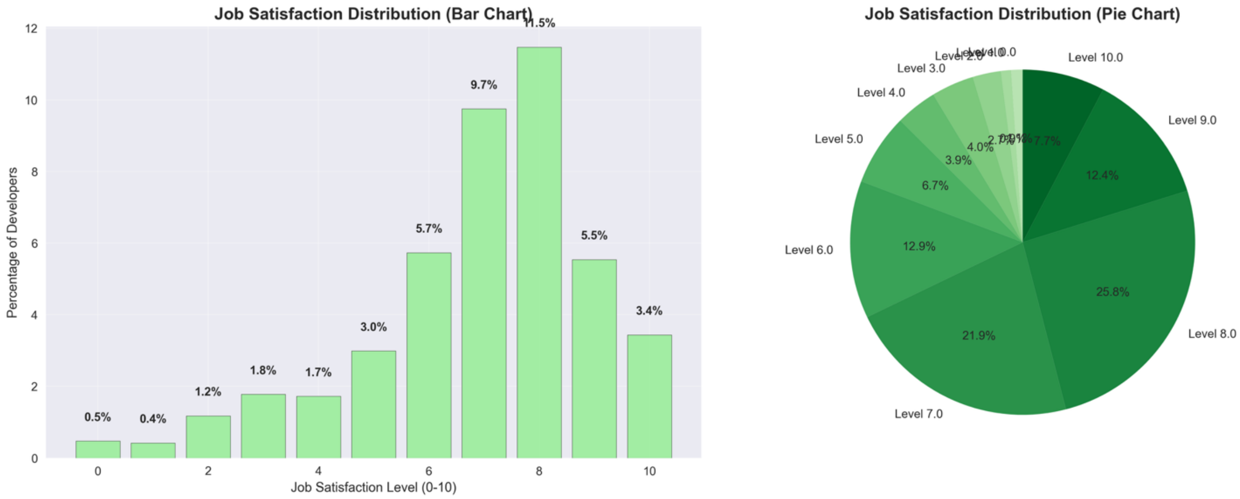


Figure 2. Compensation Distribution

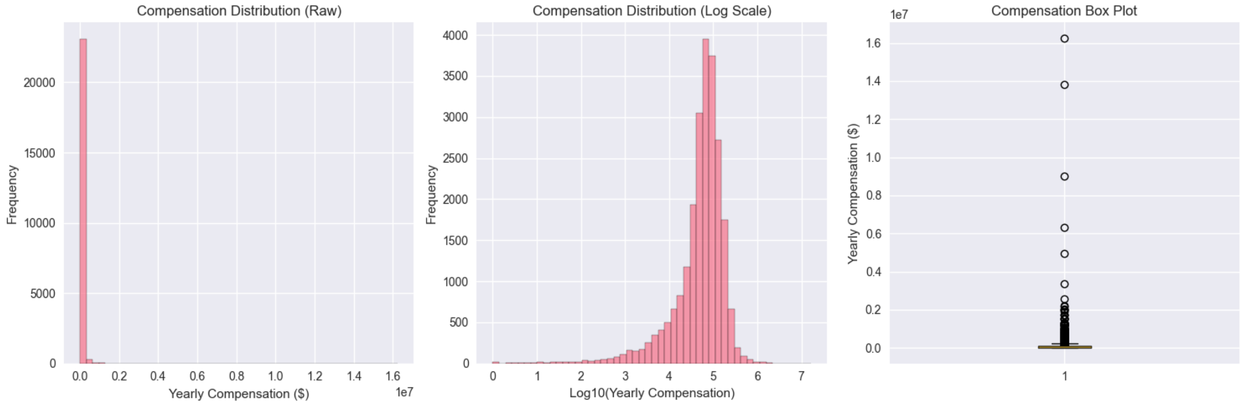


Figure 3. Top 10 Countries by Response Count

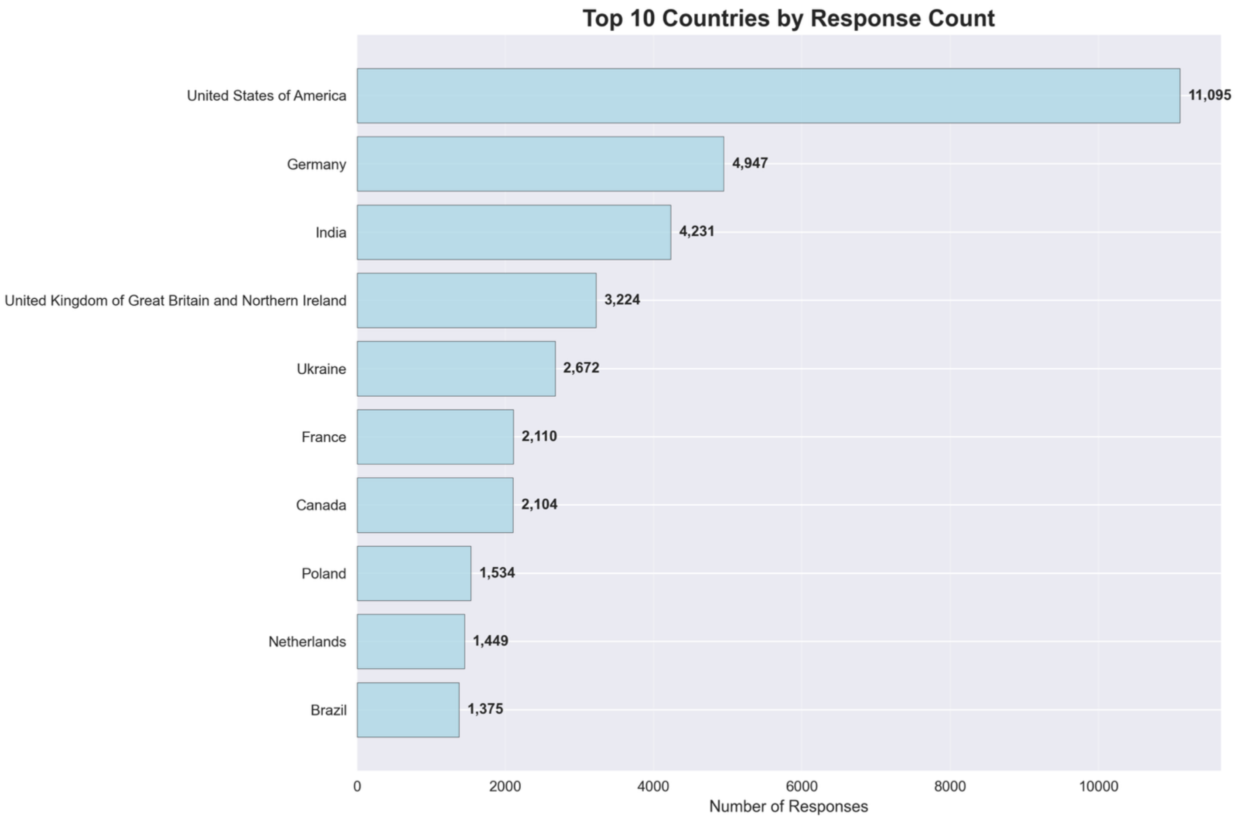


Figure 4. ROC Curves for Classification Models

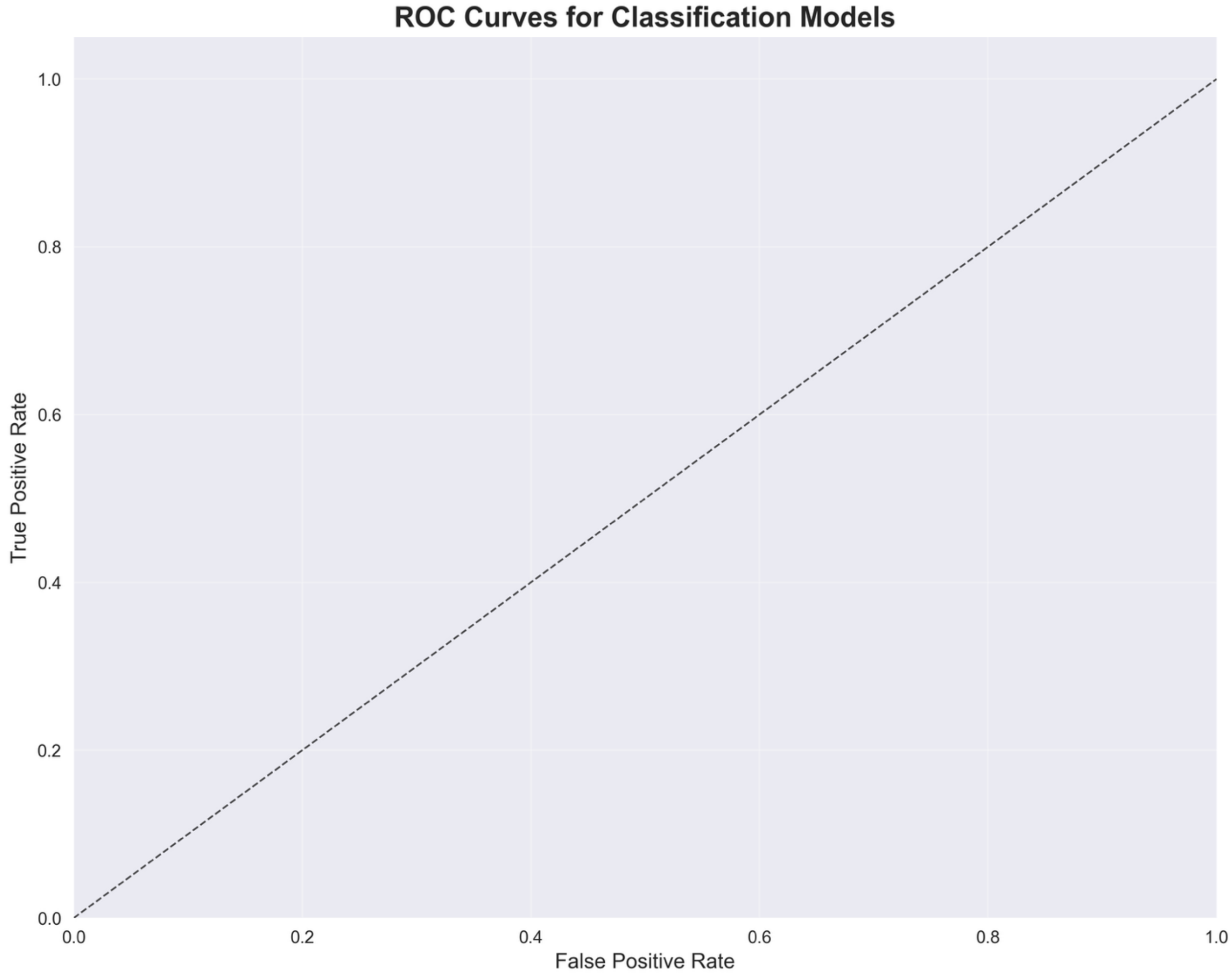
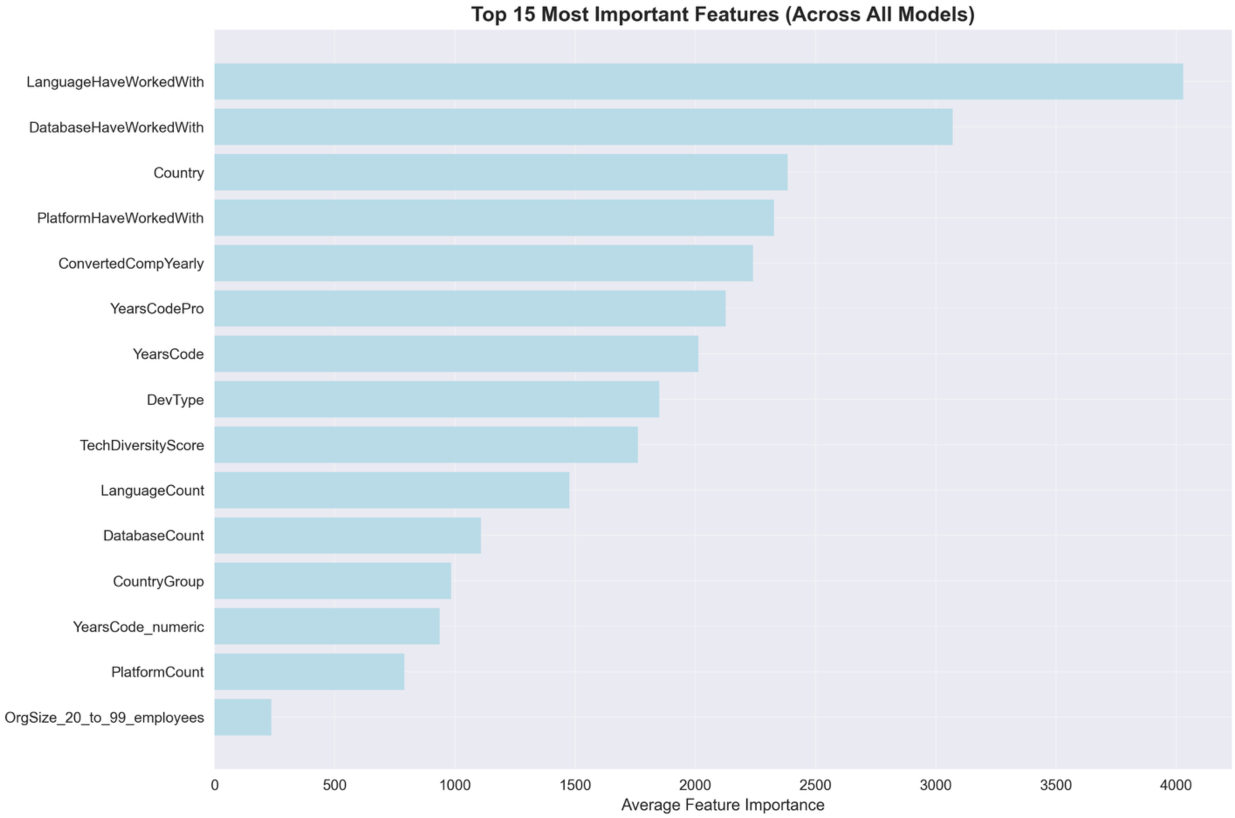


Figure 5. Random Forest Feature Importance (Classification)


Top 15 Most Important Features (Across All Models)
LanguageHaveWorkedWith
DatabaseHaveWorkedWith
Country
PlatformHaveWorkedWith
ConvertedCompYearly
YearsCodePro
YearsCode
DevType
TechDiversityScore
LanguageCount
DatabaseCount
CountryGroup
YearsCode_numeric
PlatformCount
OrgSize_20_to_99_employees
0
500
1000
1500
2000
2500
3000
3500
4000
Average Feature Importance