Universiteit Leiden

Master Computer Science

The relationship between compensation and work duration

Name: Yuchi Zhang
Student ID: s2724952

Date:

Specialisation: Computer Science: Computer Science and Science Communication and Society

1st supervisor: Niels van Weeren
2nd supervisor: Niki van Stein

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This study focuses on the critical role of compensation strategies in enhancing employee retention and satisfaction, providing essential insights for HR professionals. Utilizing data from Randstad US, an HR services industry leader, we explore the relationship between compensation and work duration across various job categories through linear regression and advanced tree-based models, including Random Forest, XGBoost, and LightGBM. Our predictive analysis shows that compensation has a notable impact on work duration, particularly in lower-level blue-collar jobs. These findings suggest that structured compensation packages can effectively extend employment tenure. Additionally, we have established a model comparison framework that systematically develops advanced classifiers with optimal predictive capabilities, demonstrating their superiority in capturing the complex interactions between compensation factors and work duration. Furthermore, we conducted group analysis to assess how predictive outcomes vary across different workforce segments, such as job levels and geographical locations, enhancing our understanding of the compensation-work duration relationship in diverse settings. Future research will expand to more sectors and incorporate additional variables to further understand the dynamics of compensation strategies.

Contents

1. Introduction

Employee retention and satisfaction are pivotal factors for the success of an organization [1], with compensation playing a crucial role in this context. Consequently, understanding the dynamics of employee compensation and its impact on work duration becomes a vital area of investigation[2].

Despite the extensive literature on compensation, the nuanced interaction between compensation and its impact on the length of employment has not been fully explored, particularly across various industries and job levels.

This project focuses on detecting the relationship between compensation and work duration, utilizing data from Randstad US, an HR services industry. Machine learning methods will be used to build assessment models and derive feature importance to find correlations between quantifiable job features.

There are two main research questions in this study:

RQ1: To what extent does compensation predict work duration?

RQ2: Does the predictability of compensation for work duration vary across different levels within the labor market?

By addressing these questions, we hope to generate valuable insights that can be utilized to improve compensation strategies, enhance employee retention, and minimize labor costs.

2. Problem Statement

The dynamics of employee compensation and its impact on work duration are complex interplays of factors crucial to workforce management and organizational effectiveness[3]. While compensation is traditionally seen as a primary motivator for employee tenure and productivity, the specifics of this relationship need to be more quantified, particularly in the context of varying job titles, levels, and industries. Understanding these specifics is important not only for developing effective compensation strategies but also for aligning these strategies with broader business purposes.

For example, in industries with high turnover rates, such as retail or hospitality, a well-structured compensation package could be strategically designed to enhance employee retention, thereby reducing recruitment and training costs. Similarly, in high-demand tech industries, where competition for talent is fierce, compensation strategies that include performance bonuses and equity packages could be crucial for attracting top-tier candidates and fostering a culture of innovation and commitment [4].

Current literature provides disparate insights into this relationship but needs a comprehensive analysis incorporating multifaceted employment characteristics. Furthermore, the increasing variability in job functions and payment structures, especially with the rise of gig and remote work, calls for a more nuanced understanding of how pay rates impact work duration.

This study addresses the gap by exploring the predictive power of compensation on work duration across different job categories and levels, employing advanced statistical models capable of delineating both linear and non-linear relationships. This study aims to answer two questions: firstly, to determine the extent to which compensation predicts work duration, and

secondly, to identify whether this predictability varies across different levels within the labor market.

By addressing these concerns, the research seeks to offer practical observations that can guide compensation strategies, improve employee retention, and streamline labor expenses. The central inquiry guiding this investigation is: "To what degree and in what way does compensation impact the length of employment across different career levels and industries?"

3. Literature Review

The literature review section aims to comprehensively analyze existing research and scholarly works concerning the relationship between compensation and work duration in the context of recruitment. This section will outline the current understanding and highlight areas that warrant further investigation by examining the relevant studies, methodologies, and findings.

Early studies in recruitment and talent acquisition relied on traditional statistical models. These models, such as linear regression and logistic regression, have provided valuable insights into the factors influencing the recruitment process [5]. However, their limitations in handling complex and nonlinear relationships have led researchers to explore the potential of machine learning algorithms [6].

Random Forests have gained popularity in recruitment research due to their ability to handle non-linearity and interactions among variables. Studies have been conducted by Frank and Moritz(2022)[7] have shown the effectiveness of decision trees in predicting compensation based on relevant factors. The research leveraging Random Forest has delved into salary predictions across more than 300 professions, utilizing a dataset encompassing over three million employees. This approach stands out for learning distinct random-forest models for each profession, thereby accommodating the specificities of each field. The findings from this study are meaningful, demonstrating a mean absolute percentage error (MAPE) of 17.1% in salary predictions, surpassing previous machine-learning benchmarks. A vital attribute of this method is its capacity to manage categorical variables effectively, reducing their cardinality, and its adeptness in outlier detection and handling. By training separate models for each profession, the Random

Forest approach accounts for the heterogeneity of the salary determinants, leading to a nuanced and accurate predictive model. This methodology enhances the quality of salary predictions and provides insights into the factors most influential in determining compensation. This knowledge is valuable for organizational planning and employee retention strategies.

The related research by Rohit Punnoose(2016)[8] discusses the problem of employee turnover in organizations and the use of machine learning algorithms, specifically Extreme Gradient Boosting (XGBoost), to predict it. The article compares the performance of XGBoost against six other supervised classifiers on a dataset of 73,115 data points from a global retailer's HRIS database and finds that XGBoost outperforms the other classifiers in terms of accuracy, runtime, and memory utilization.

According to the dataset we utilized, we found that Randstad's employer branding experts attract a lot of gig economy talent[9]. Temporary and part-time positions often get filled relatively quickly; however, these roles also tend to have shorter durations of employment, indicating a higher turnover rate. The gig economy is an increasingly important concept that has far-reaching implications for workers and employers in most sectors[10]. The Bureau of Labor Statistics [11] notes that, while there is no single or official definition of the gig economy, a 'gig' generally refers to "a single project or task for which a worker is hired, often through a digital marketplace, to work on demand". Other definitions of the concept have emphasized how "temporary, flexible jobs are commonplace" in the gig economy and companies "tend to hire independent contractors and freelancers instead of full-time employees".

In summary, early research work in related fields mainly used linear models; the subsequent use of machine learning models capable of recognizing more complex nonlinear patterns, such as random forests and gradient boosted trees, which resulted in good performance, inspired the choice of models in this research. However, previous research lacked a refined dis-

cussion of the different levels of the labor market on the one hand and did not take today's gig economy into account on the other.

The dataset used in this paper is dominated by temporary jobs, which can well study the relationship between salary and work duration in this new gig economy, and at the same time, through the discussion of refinement at different levels of the labor force, it can give more practically meaningful findings and conclusions.

4. Data collection and construction

In this chapter, we detail the process of data collection and preparation for building models. We describe the data sources, the features extracted from the raw data, and the preprocessing steps taken to clean and transform the data. This groundwork is crucial to ensure the quality and relevance of the dataset for subsequent machine learning modeling.

4.1 Data Source

This research used data from Randstad USA[12], a major workforce solutions provider, through Google BigQuery [13], a fully managed data warehouse for large-scale analytics. This dataset comprises an extensive collection of historical assignment data over the past years, encapsulating millions of rows and hundreds of columns. The data encompasses a wide range of variables, including but not limited to pay rate, job title, job level, career area, state, and work duration, as well as other assignment-specific information. As a result of the dataset's size and diversity of features, it serves as a valuable resource for studying compensation across different dimensions of the US labour market. See Appendix A.1 for specific fields in the raw data.

4.2 Data Features

The historical data used to build the models contains the following features:

- **Pay Rate:** represents the amount of money a worker is paid per hour
- **Job Title:** represents the job title of the assignment
- **Job Level:** the experience level associated with the job, such as "Entry Level."

- **Career Area:** refers to a specific field or industry in which an individual pursues their professional career, such as "Manufacturing and Production."
- **State:** represent the geographical information of the location where the assignment takes place in the US
- **Work Duration:** represents the number of days into the assignment
- **Business Unit:** data is derived from multiple business sources such as RPEUS, RNALP. Our data mostly comes from the RNALP business unit. RNALP involves more blue-collar work, with relatively short work durations, similar to temporary work. On the other hand, units such as RTDUS, RPEUS, RPUUS, and RPOUS are focused on technical white-collar work.

4.3 Data preprocessing

The process of preparing collected US pay rate data involves a series of steps that transform raw data into a format that is easy to understand. This involves cleaning, filtering, labeling, enriching the data, and conducting other feature engineering steps to ensure the data's quality meets the standards required for model development. The data preprocessing includes these steps:

- **Filter assignments based on date**
  The model for US pay rates is constructed using datasets that include the historical pay rates data from 2018-09 to 2023-09; therefore, data older than five years are filtered out.
- **Remove redundancy**
  As part of the dataset, the RNA_REPORT_DATE column provides a detailed record of pay rate adjustments for each assignment. The source database is updated weekly to reflect the employee's settled pay rate, resulting in numerous duplicate values for the same assignment. RNA_REPORT_DATE represents the date on which each compensation is inserted into the database. To remove redundancy, we

only retain the initial and most recent pay rates for each assignment. If the initial and final pay rates differ, indicating a compensation adjustment, the final pay rate is calculated as the average of these two values.

- **Filter assignments based on Burning Glass job title taxonomy score**
  In this phase, each assignment is aligned with a normalized job title according to the Burning Glass taxonomy[14]and a matching confidence score represents the accuracy of this alignment. The method for normalizing job titles involves using tables stored in the BigQuery database. Initially, job titles are standardized by converting them to lowercase and removing spaces and special characters. The process utilizes two main tables, one contains normalized job titles, and one includes detailed job levels and scores. Using SQL queries, job titles from assignments are matched to the normalized titles in the database, ensuring consistency and accuracy. The score represents how accurately the job assignment matches a specific job title within the taxonomy. The score ranges from 0 to 1, with higher scores indicating a better match. Assignments with a confidence score below 0.75 are filtered out and not included in the following steps, ensuring that only those with a high degree of mapping certainty are considered.
- **Filter pay rates based on minimum wage and maximum cutoff**
  Bill rates lower than $7.25 USD (US Federal minimum wage) or higher than $350 USD are excluded from the next steps.
- **Drop unnecessary columns**
  The model is trained using the following features: Pay Rate, Job Title, Job level, Career Area, State, Work Duration, and Business Unit. The remaining features are discarded.
- **Split dataset into train and test sets**
  The raw dataset is shuffled and split into two subsets before building a model: the training set and the test set. The training set, which constitutes 80% of the initial dataset, is used for developing the model, while the testing set accounts for the remaining 20% and is utilized for

external evaluation.

- **Data Transformations**

  To enhance model performance, we apply certain transformations to the entire dataset. The initial step involves converting categorical features to numeric values, enabling their inclusion in regression models. This is achieved by transforming all non-numeric labels into numerical ones using the LabelEncoder class from sklearn[15].

  Secondly, we standardize numeric features to ensure uniformity in scale, a step critical for algorithms that perform better with scaled data. It is important to note that this standardization primarily affects the linear regression model. Tree-based models, such as Random Forest and XGBoost, do not require standardized features because they are insensitive to the scale of the data due to their splitting criteria. In contrast, linear regression models are sensitive to feature scales, and standardized data helps improve their performance and convergence.

  This standardization, achieved through StandardScaler from sklearn, normalizes the data to have a mean of 0 and a standard deviation of 1, assuming the data follows a normal distribution.

With the data collection, feature extraction, and preprocessing steps completed, the dataset is now ready for machine learning modeling.

5. Methods

In this chapter, we describe the various machine learning techniques selected for predicting employee work duration.

We selected a linear regression model and three tree-based models, based on our literature research, which highlights their effectiveness in handling similar predictive tasks. We aim to provide a comprehensive approach by including both linear and tree-based models, allowing us to capture a wide range of data characteristics and improve prediction accuracy.

These models are well-suited for this task for several reasons. Linear regression is chosen for its simplicity and interpretability, making it easy to understand the relationship between the dependent and independent variables. Its straightforward nature allows us to quickly identify and quantify the impact of different features on work duration.

On the other hand, tree-based models such as Random Forest, XGBoost, and LightGBM are included due to their capability to handle complex and non-linear relationships within the data. These tree-based models represent the main types of ensemble learning methods and are among the most advanced techniques currently available. Random Forest is known for its robustness and ability to reduce overfitting by averaging multiple decision trees. XGBoost, or eXtreme Gradient Boosting, is highly efficient and effective, particularly due to its optimization techniques and ability to handle large datasets. LightGBM, or Light Gradient Boosting Machine, enhances speed and performance through its unique approach to tree-based learning, making it suitable for high-dimensional data.

By leveraging these advanced methods, we can better capture the intricate

patterns and interactions within the dataset, leading to more accurate and reliable predictions of employee work duration.

Next, the theoretical details of these methods are presented, including machine learning models, feature importance measures, and statistical testing methods.

5.1 Linear Regression

Linear regression [16] is a foundational statistical technique used to predict the value of a dependent variable based on the value of one or more independent variables. In the context of this study, linear regression is used to model the dependence of WorkDuration (Y) on the explanatory variable PayRate (X) along with other covariates.

The model is expressed by the following equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + \epsilon_i$$

Where: Y_i is the dependent variable representing the work duration of the i^{th} individual,
$X_{1i}, X_{2i}, \ldots, X_{ni}$ are the independent variables for the i^{th} individual,
β_0 is the y-intercept,
$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for each independent variable,
ϵ_i is the error term for the i^{th} individual.

The coefficient β_j associated with each independent variable X_j indicates the amount of change one can expect in WorkDuration given a one-unit change in X_j, assuming that other variables are kept constant.

The significance of each β_j is assessed by its p value, obtained from the

t-test:

$$t = \frac{\beta_j}{\text{SE}\left(\beta_j\right)}$$

where $SE\left(\beta_j\right)$ denotes the standard error of the coefficient β_j. A p-value less than 0.05 typically indicates statistical significance.

The goodness of fit for the model is evaluated by the R^2 statistic, which is the proportion of variance in the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - Y\right)^2}$$

where $\hat{Y}_i$ is the predicted value of WorkDuration and $\bar{Y}$ is the mean value of WorkDuration.

In our study, the linear regression model's $\left(R^2\right)$ value suggests that additional variables and model complexity may be required to fully explain the variance in work duration, prompting tree-based models for further analysis.

5.2 Random Forest Regression

The Random Forest Regression algorithm [17] is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. This method enhances predictive accuracy and controls over-fitting. The model's feature importance is assessed by the extent to which each feature decreases the impurity of the nodes, often measured by the reduction in variance.

The importance of a feature can be quantified as the sum over the number

of splits that include the feature, proportionally to the number of samples it splits:

$$\text{Importance}\left(X_m\right) = \sum_{t \in T} p(t) \cdot \Delta i\left(s_t, t\right)$$

where T is the set of trees, t indexes the trees, $p(t)$ is the proportion of samples that reach node t, s_t is the splitting criterion at node t, and $\Delta i\left(s_t, t\right)$ is the impurity decrease resulting from that split.

5.3 XGBoost Regression

XGBoost stands for eXtreme Gradient Boosting[18] and is a sophisticated algorithm that solves regression and classification tasks using an advanced version of Gradient Boosted Decision Trees. Fundamentally, XGBoost is an ensemble approach that leverages the concept of gradient boosting by transforming a group of basic models into a single, more robust model. The essence of boosting involves building new models that address the shortcomings of previous ones, which results in incremental improvements.

Gradient boosting utilizes gradient descent to systematically reduce errors between models in a sequential manner. This approach emphasizes minimizing the loss function, which is a user-customizable error metric, thus enabling more precise control over the model's optimization process.

XGBoost is an efficient and advanced implementation of the Gradient Boosting Decision Tree framework, which stands out as one of the most favored models in gradient boosting. Its good performance is attributed to several innovative features:

- It utilizes a variety of optimization techniques in software and hardware, achieving speeds up to ten times faster than a gradient boosting machine;
- The algorithm's scalability is heightened by its ability to construct

trees in parallel;

- It opts for a depth-first approach to tree pruning, resulting in faster convergence compared to breadth-first strategies;
- Improves model generalization by penalizing complex models with L1 (LASSO) and L2 (Ridge) regularisation terms during training;
- It learns the tree branch directions for missing data during training to handle missing data values more efficiently.

In the ensemble method used by XGBoost, it constructs a sequence of K decision trees, denoted as (f_k). Each tree in the sequence is designed to correct the errors made by the tree before it. The predictive model is expressed as:

$$y_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where y_i is the prediction for the i-th training sample, $\mathcal{F}$ is the space of functions (trees), and K represents the number of trees. The training objective combines a differentiable loss function $l(y_i, \hat{y}_i)$ and a regularization term $\Omega(f_k)$:

$$\text{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $\Omega(f_k)$ penalizes the complexity, ensuring the simplicity of the model.

5.4 LightGBM Regression

LightGBM (Light Gradient Boosting Machine)[19] is a gradient boosting framework that uses tree-based learning algorithms and is designed for speed and efficiency. It is particularly effective for large datasets and high-

dimensional data. Similar to XGBoost, LightGBM computes feature importance by counting the number of times a feature is used in model construction. However, LightGBM improves upon the traditional methods by using histogram-based algorithms, which bucket continuous feature values into discrete bins to speed up the training process.

The feature importance metric for LightGBM can be formalized as follows:

$$\text{Importance}\left(X_m\right) = \sum_{j=1}^{J} I\left(f_j = X_m\right)$$

where J is the number of splits across all trees in the model, f_j denotes the feature used for the j th split, and I is the indicator function.

5.5 Likelihood Ratio Test

The likelihood ratio test (LRT) [20] is a statistical procedure that tests the goodness of fit between two competing statistical models based on the ratio of their likelihood functions. Specifically, the LRT evaluates the null hypothesis, H_0, which posits that a simpler model provides an adequate fit to the data, against an alternative hypothesis, H_1, which suggests that a more complex model is necessary. The test statistic is derived by calculating Λ, the ratio of the maximum likelihood of the data under the null hypothesis ($\mathcal{L}(\theta_0)$) to that under the alternative hypothesis ($\mathcal{L}(\theta_1)$), expressed as:

$$\Lambda = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_1)}$$

Under the null hypothesis, and assuming regularity conditions, $-2\log(\Lambda)$ asymptotically follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters estimated by the two models. This asymptotic property allows researchers to compute p-values

and thus assess the statistical significance of the difference in fit between the two models.

In this paper, the LRT was utilized to compare two models: one that included the wage variable (PayRate) and one that did not. By comparing these two models, we aim to determine whether including the wage variable improves the model's fit for job duration prediction.

5.6 Paired t-test

The paired t-test [21] is a statistical procedure used to determine whether the mean differences between two sets of observations are statistically significant. This test is applicable when the data points are paired, meaning that each measurement in one group is uniquely linked to a specific measurement in the other group. This is typically the case where we need to compare two sets of related data to determine if there is a statistically significant difference between them.

The test statistic for the paired t-test is calculated as follows:

$$t = \frac{\overline{d}}{s_d / \sqrt{n}}$$

where $\overline{d}$ is the mean of the differences between paired observations, s_d is the standard deviation of these differences, and n is the number of pairs.

The null hypothesis H_0, asserts that the mean difference is zero, while the alternative hypothesis, H_1, claims that the mean difference is not zero. Under the assumption that the differences are normally distributed, the test statistic follows a t-distribution with $n - 1$ degrees of freedom. This allows for the calculation of a p-value to assess the significance of the observed difference.

Specifically, we use paired t-tests to compare the prediction results of different models (e.g., linear regression and tree-based models such as Random Forest, XGBoost, LightGBM) with and without the inclusion of the wage variable (PayRate). This comparison is mainly used to test whether the change in model prediction performance is significant, i.e., whether the inclusion of the wage variable improves the prediction of the model.

6. Experimental design

The empirical investigation of our study is designed to systematically evaluate the predictive performance of various regression models and to understand the impact of PayRate on the duration of work assignments, as explored through both linear and non-linear methodologies.

By modelling the relationship between PayRate and the duration of work assignments through both linear and non-linear methodologies, we are able to provide quantitative evidence for the relationship between PayRate and the duration of work assignments, as well as explore whether there are differences in this relationship across different labour market subgroups.

6.1 Evaluation Methods

The model evaluation is based on several statistical metrics. Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. Mean Squared Error (MSE) is similar but squares the differences before averaging to penalize larger errors. Root Mean Squared Error (RMSE) takes the square root of MSE, thus providing error metrics in the same unit as the original data. The Coefficient of Determination, denoted as R^2, quantifies the amount of variance in the dependent variable that is predictable from the independent variables [22].

To compare the performance of models with and without PayRate, we apply the likelihood ratio test [23], which assesses whether the inclusion of PayRate significantly improves the model fit. We also conducted a comparison of model performances across different business units to assess the outcomes under various business unit conditions. Additionally, we employ a paired t-test to statistically examine the difference in predictions from the

two sets of models.

6.2 Linear Regression Exploratory Analysis

A linear regression model is initially applied to predict WorkDuration, utilizing PayRate as an explanatory variable. The dataset is also grouped by Business Unit and Job Level to explore differences between them. By examining the regression coefficients, we can assess the strength and nature of the relationship between PayRate (X) and WorkDuration (Y).

6.3 Regression with Tree-based Models

Building upon the linear analysis, we employ tree-based models such as Random Forest[24][25], XGBoost [26], and LightGBM [27]. These models excel at unraveling complex, non-linear relationships and interactions among predictors. Assessing feature importance becomes a valuable tool for comprehending each feature's contribution to work duration, allowing us to analyze the significance of payload in determining work duration (where 'X' represents all features and 'Y' represents work duration).

The model development phase involves a series of critical steps: training the model [28], tuning hyperparameters [29], and validating model performance.

Training regression model

The training process for the different tree-based regression models is similar, preparing the training data, using the squared loss function. Models learn to minimize errors by recursively partitioning the feature space. Iterative node splitting guides the creation of decision rules [30].

Hyper-parameter tuning

Hyperparameter tuning [31] is a crucial aspect of model development, particularly to enhance model accuracy and predictive power. We employ the 'GridSearchCV' method [32] from the 'sklearn' library [33] to systematically explore a range of hyperparameter combinations and identify the optimal settings for each model, as it is the simplest in principle and implementation and the most comprehensive in candidate parameter coverage, suitable for situations where the data volume and the set of candidate parameters are not large. 'GridSearchCV' conducts an exhaustive search over specified parameter values for an estimator. The parameters of the estimator used to apply these methods are optimized by cross-validated grid-search over a parameter grid [34]. Appendix A.2 provides an overview of the hyperparameters explored for the three different models under investigation.

Model Performance Validation

Validation is conducted through external validation [35]. A dataset that has not been previously exposed to the model is used to further validate the model's predictive capability after it is trained. This step is crucial for assessing the model's generalizability to new data. On this dataset, predictions are evaluated, and the R^2 value is calculated to compare model performance. This comparison between internal and external validation metrics helps identify any overfitting or underfitting issues. Additionally, error metrics such as the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are calculated, providing a comprehensive view of the model's predictive accuracy [36].

If a model that relies on payrate as its main feature is able to successfully predict work duration on unseen data, this means that the model is truly discovering the patterns in it and how payrate affects, or even determines, work duration.

6.4 Model Comparison In Pairs

To quantify the impact of PayRate, we perform comparative regression analyses using tree-based models in 6.3 both with and without this feature. By excluding PayRate, we can observe any changes in predictive performance, as indicated by our evaluation metrics (MAE, MSE, RMSE, R^2). The LRT [37] provides a statistical basis to assess whether PayRate's inclusion enhances the model fit significantly. The paired t-test is applied to compare the mean prediction errors of models with and without PayRate, thereby offering a paired comparison of model performances [38].

To ensure the validity and repeatability of our findings, we replicate the model comparison process ten times [39]. This replication aids in accounting for variability and enhances the statistical power of the tests. The results from these repeated experiments will be aggregated to furnish a comprehensive understanding of PayRate's significance.

7. Exploratory Data Analysis

This chapter presents the results of the experiment based on linear regression analysis. We generated visualizations that show the relationship between pay rate and work duration.

7.1 Basic Visualization

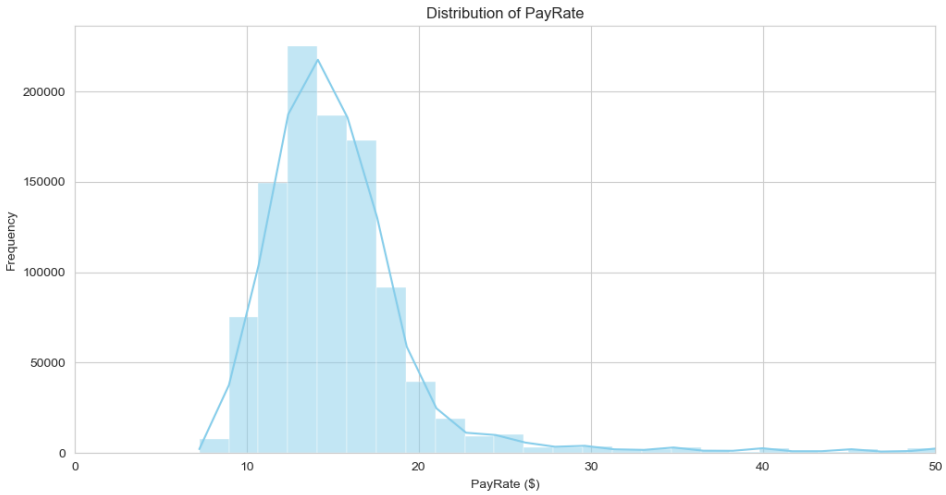
Figure 7.1 visualizes the frequency distribution of PayRate and WorkDuration.

It can be noticed that most of the samples have a pay rate concentrated in the 10-20 range and work duration concentrated in less than 30 days, which matches well with the source of our dataset: predominantly short-term, blue-collar jobs.

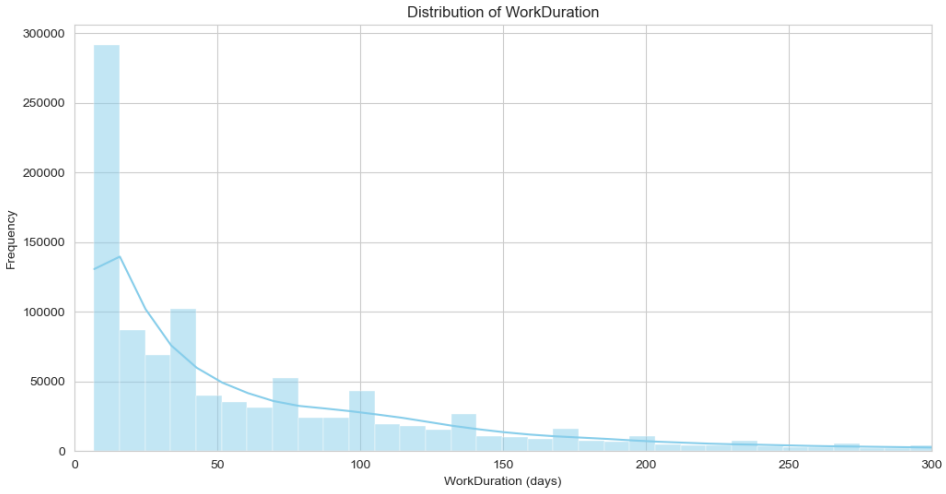
7.2 Linear Regression Exploratory Analysis

7.2.1 OLS Regression Results

Figure 7.2 (a) depicts a regression plot that illustrates the relationship between PayRate and Work Duration, it suggests a positive correlation between PayRate and Work Duration.
Figure 7.2(b) displays the results of the Ordinary Least Squares (OLS) regression analysis: The coefficient for PayRate (x1) is **0.2341**. Since both variables have been standardized, this means that for every standard deviation in rease in PayRate, the WorkDuration increases by about 0.2341 standard deviations.
The t-value for the PayRate coefficient is **245.309**, and its associated p-value $(P > |t|)$ is **0.000**, which indicates that PayRate is a statistically significant predictor of WorkDuration.



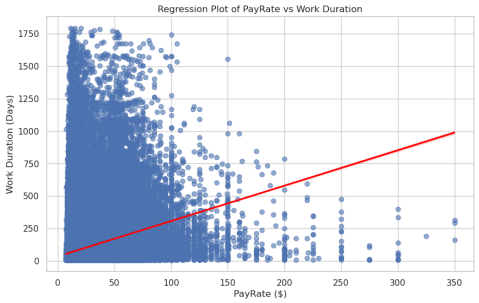
(a) Distribution of PayRate



(b) Distribution of WorkDuration

Figure 7.1: Distribution of PayRate and WorkDuration

The R squared value is **0.055**, which means that PayRate explains about 5.5% of the variance in WorkDuration, suggesting that although there is a relationship, many other factors not included in the model likely influence WorkDuration.



(a) Regression Plot of Payrate VS Work Duration

OLS Regression Results

Dep. Variable:	y	R-squared:	0.055
Model:	OLS	Adj. R-squared:	0.055
Method:	Least Squares	F-statistic:	6.018e+04
Date:	Sat, 18 Nov 2023	Prob (F-statistic):	0.00
Time:	23:25:46	Log-Likelihood:	-1.4429e+06
No. Observations:	1037489	AIC:	2.886e+06
Df Residuals:	1037487	BIC:	2.886e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.433e-17	0.001	7.79e-14	1.000	-0.002	0.002
x1	0.2341	0.001	245.309	0.000	0.232	0.236

Omnibus:	906854.159	Durbin-Watson:	1.845
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40811546.957
Skew:	4.072	Prob(JB):	0.00
Kurtosis:	32.627	Cond. No.	1.00

(b) OLS Regression Result

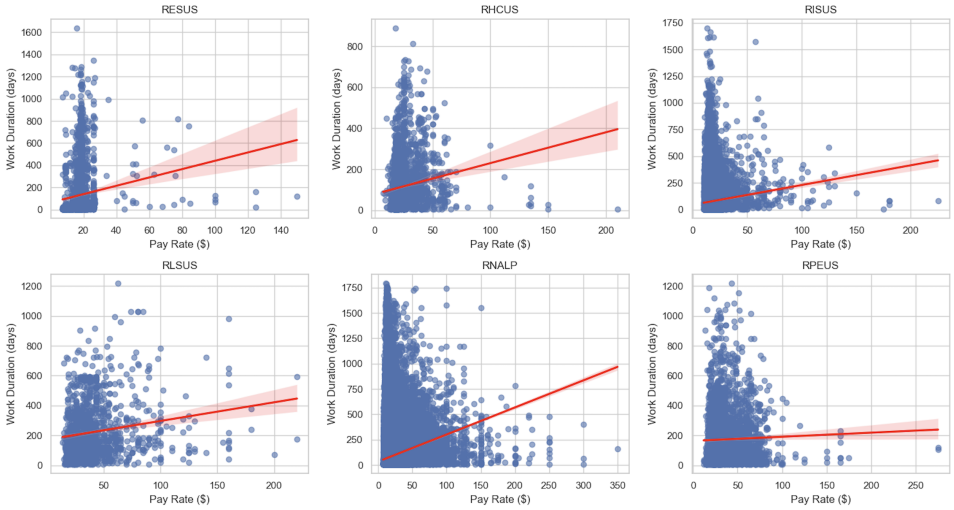
Figure 7.2: Comparative Analysis of Payrate and Work Duration with Ordinary Least Squares Regression

We further analysis the correlation between PayRate and WorkDuration across different Business Units and States. As depicted in Figure 7.3 and Figure 7.4, we present a subset of these regression plots, selected to demonstrate the variation in relationships across the categories. Due to the extensive number of plots, only representative samples are displayed.

7.2.2 Regression Analysis Across Business Units

As we can see in Figure 7.3, the coefficient values (coef) exhibit noticeable variation across different business units, ranging from 0.274322 (RPEUS) to 3.914674 (FRNUS). This indicates a significant variability in how PayRate impacts WorkDuration across different business units. For instance, in FRNUS, each additional dollar in PayRate is associated with an increase in WorkDuration by approximately (3.91). Conversely, in RPEUS, WorkDuration increases by only approximately (0.27) for each additional dollar in PayRate.

The (R^2) values are generally low across all business units, underscoring that PayRate explains a small portion of the variation in WorkDuration within



(a) Regression plot by BusinessUnit

BusinessUnit	coef	intercept	r_squared	p_value
FRNUS	3.914674	44.822588	0.014441	4.49E-35
PICUS	0.837583	200.421499	0.00865	3.04E-13
RESUS	3.754183	64.752889	0.016071	9.05E-12
RHCUS	1.505384	79.400483	0.021484	2.92E-15
RISUS	1.847976	45.522619	0.006055	6.51E-110
RLSUS	1.249895	171.730345	0.028583	4.14E-10
RNALP	2.669869	33.452889	0.020547	0.00E+00
RPEUS	0.274322	163.574618	0.000859	9.57E-02
RPUUS	1.520077	74.106161	0.027065	4.66E-96
RTDUS	0.327866	190.332665	0.000996	1.78E-08
SPHUS	1.895257	40.056004	0.007055	1.01E-250

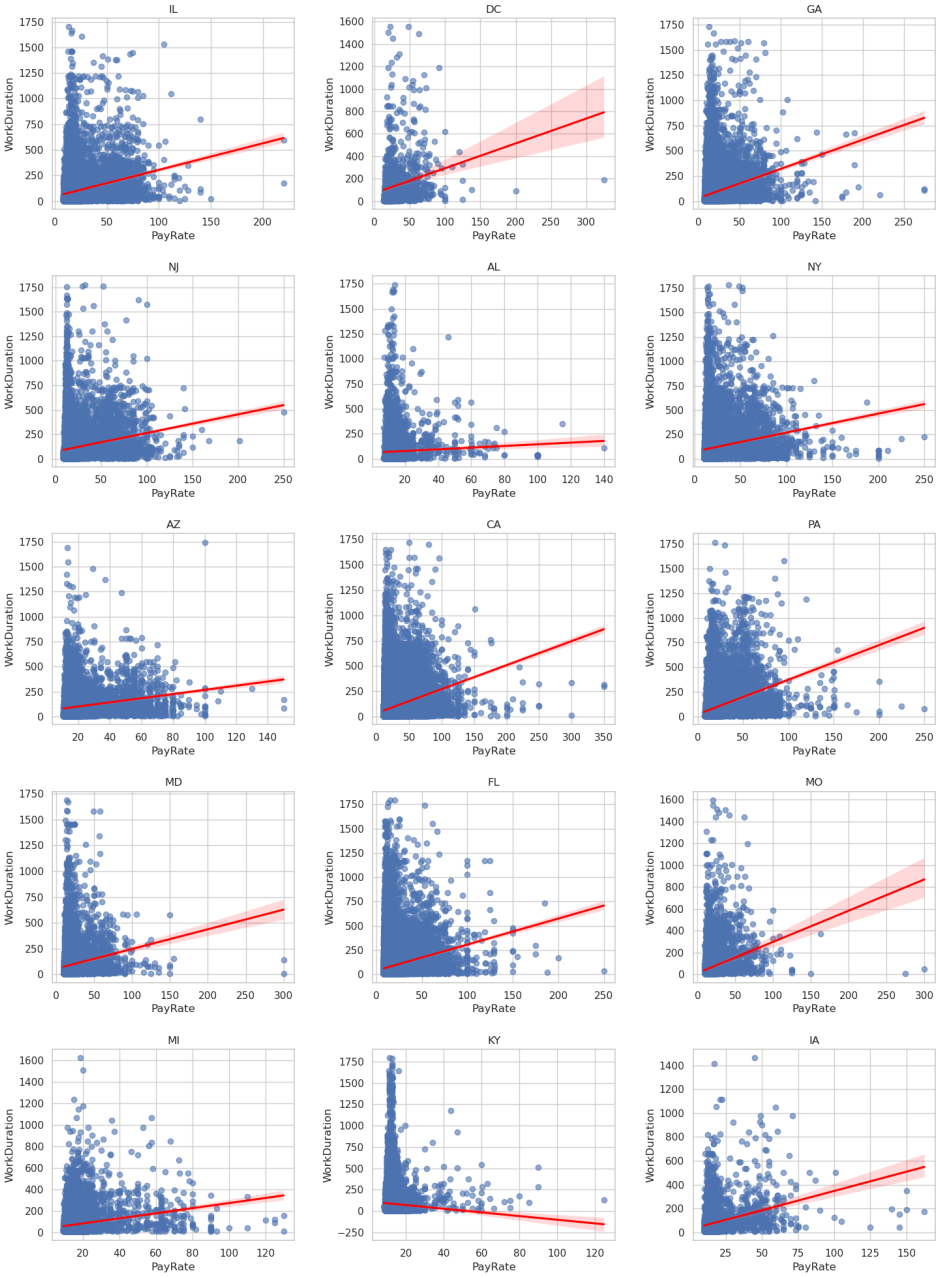
(b) Regression Analysis Coefficients and Statistics by Business Unit

Figure 7.3: Comparative Regression Analysis Across Business Units

each business unit. The highest$\left(R^2\right)$ is observed in RLSUS (0.028583), yet even this value is quite low. This implies that other factors, in addition to PayRate, contribute to the variability in WorkDuration.
The p-values also exhibit variation across different business units. Certain business units like RNALP display extremely low p-values, which means that PayRate is a statistically significant predictor of WorkDuration for these units. On the other hand, business units like RPEUS and RTDUS show high p-values, suggesting that PayRate does not statistically significantly predict WorkDuration in these contexts.

7.2.3 Regression Analysis Across States

According to Figure 7.4, the coefficients showcase substantial variation across states, ranging from a high of 3.69 in WV to a low of (-2.14) in KY. This extensive range underscores the diverse impact of PayRate on WorkDuration across different states. Most states exhibit a positive relationship, where a higher PayRate correlates with increased WorkDuration, while a few like KY indicate a negative correlation.
The $\left(R^2\right)$ values generally lean towards the lower end, demonstrating that PayRate accounts for a minimal portion of the variability in WorkDuration across most states.
The majority of the states present low p-values, highlighting the statistical significance of PayRate in predicting WorkDuration in these states. However, states like HI, AK, MS, WY,and VT display higher p-values, suggesting the absence of a statistically significant relationship between PayRate and WorkDuration in these regions.



(a) Regression plot by State, the unit of the x-axis is $, and the unit of the y-axis is days

State	coef	intercept	r_squared	p_value
HI	1.526953	180.846331	0.00734	0.629978
AK	-1.004243	140.011288	0.014641	0.469287
MS	0.312002	70.853469	0.000154	0.368689
WY	2.835616	70.118277	0.024449	0.34182
VT	0.862615	88.549576	0.000892	0.339461
SD	1.026773	75.250512	0.004103	0.179866
ME	1.921204	121.125719	0.045244	0.000541
ID	2.29952	56.085257	0.027558	0.00032
ND	1.262172	68.023289	0.017663	0.000194
AL	0.837451	62.420701	0.001489	0.000013
WV	3.692405	14.768964	0.022237	0
FL	2.675431	39.863453	0.02945	0
TX	2.571357	39.39093	0.086936	0
UT	2.558971	33.793444	0.012765	0
MI	2.367463	35.15197	0.030203	0
AR	2.314042	35.514746	0.024647	0
DC	2.214852	72.376291	0.044515	0
NY	1.941485	75.567557	0.028377	0
NV	1.722734	44.639152	0.019606	0
KY	-2.143442	112.123919	0.003703	0

(b) Regression Analysis Coefficients and Statistics by State

Figure 7.4: Comparative Regression Analysis Across States

In this chapter, we explored the relationship between pay rates and work duration using linear regression analysis. We provided a series of visualizations to illustrate the distribution of pay rates and work durations, and confirmed their positive correlation through Ordinary Least Squares (OLS) regression analysis. The findings indicate that higher pay rates are associated with longer work durations, which is essential for understanding the mechanisms of employee motivation. However, the explanatory power of the model is limited, with an R-squared value of 0.055, indicating that there are likely many other influencing factors not included in our model. Moreover, the regression analyses across different business units and states in the United States further highlight the complexity and variability of how pay impacts work duration, revealing differences across various levels of the labor market. Overall, the analysis in this chapter, while revealing, is not exhaustive. It not only sheds light on the relationship between pay and work duration but also emphasizes the need for further multidimensional and subgroup analyses to fully understand how compensation strategies impact employee retention across different industries and regions.

8. Machine Learning Prediction Models

This chapter begins with an ablation experiment: whether or not to include payrate in the features is used as a control variable to construct two sets of machine learning models, and then performance metrics and statistical tests (likelihood ratio tests, paired t-tests) are used to compare the predictive results of the predictive models with and without the inclusion of payrate. The machine learning methods used include linear regression, random forest, XGBoost, LightGBM.

Then, the importance of the features of the machine learning model is calculated to view the important contribution of the payrate variable to the duration of the prediction effort from another aspect.

Finally, to ensure that we have constructed the best prediction model and to exhaustively compare the performance of different machine learning methods on this task, a hyperparametric search of the three tree-based prediction models is performed to compare their prediction performance on the optimal parameters.

8.1 Ablation experiment: PayRate

The models with PayRate as a feature show better performance in predicting work duration compared to the models without PayRate. As can be seen from the lower RMSE and MAE values, indicating more accurate predictions with fewer errors. The (R^2) value, which explains the variance in work duration by the model, is also slightly higher when PayRate is included, suggesting that PayRate contributes to the model's explanatory power.

The results from the analysis using Random Forest, XGBoost, and Light-

	RMSE	MAE	R^2	MSE
w/ PayRate	108.131313	66.857335	0.061810	11692.381055
w/o PayRate	109.853304	68.232725	0.031690	12067.748580

Table 8.1: Metrics for Linear Regression w/ and w/o PayRate

	RMSE	MAE	R^2	MSE
w/ PayRate	100.477855	60.379816	0.189919	10095.799263
w/o PayRate	104.733945	64.420022	0.119838	10969.199209

Table 8.2: Metrics for Random Forest w/ and w/o PayRate

	RMSE	MAE	R^2	MSE
w/ PayRate	102.503274	63.232141	0.156931	10506.921081
w/o PayRate	104.232874	64.424056	0.128239	10864.492048

Table 8.3: Metrics for XGBoost w/ and w/o PayRate

	RMSE	MAE	R^2	MSE
w/ PayRate	103.084581	63.909881	0.147341	10626.430895
w/o PayRate	104.710097	64.847587	0.120238	10964.204431

Table 8.4: Metrics for LightGBM w/ and w/o PayRate

GBM models indicate an association between compensation and work duration. The P-values from the Likelihood Ratio Test and Paired t-test suggest that compensation is a reliable predictor of work duration. In these tests, the models' good predictive capability can be seen by their extremely low P-values.

Table 8.5: Statistical Test Results for Machine Learning Models

Model	Likelihood Ratio Test P-value	Paired t-test P-value
LightGBM	1.19×10^{-138}	4.22×10^{-13}
XGBoost	7.06×10^{-228}	5.62×10^{-12}
Random Forest	0.0	4.63×10^{-12}
Linear Regression	7.98×10^{-183}	7.45×10^{-15}

In fact, the linear regression model is a very poor fit in terms of R^2, which led us to abandon its further analysis, because if the fit is poor, then the importance of the features calculated from it may not be meaningful. Since linear regression has no hyperparameters to adjust, it was abandoned in the later modeling analysis, and only the three tree-based models were subsequently analyzed.

8.2 Feature importance

The feature importance plots, as shown in Figure 8.1, confirm the relevance of PayRate as it shows the highest importance score, followed by other features. PayRate's dominance underscores its value as a determinant of work duration.

In addition to this, the ordering of important features is identical for all three models, but Payrate is more important in Random Forest and its prediction is the best. This leads one to speculate that Random Forest learned more effective prediction patterns about Payrate.

8.3 Hyper-parameters Tune

In order to further explore the optimal prediction performance of the model, the optimal parameters are found by cross-validation grid parameter search for Random Forest, XGBoost and LightGBM [40].See Appendix A.2 for candidate parameters.

The optimal parameters are shown in Table 8.6.

Model	Best Parameters

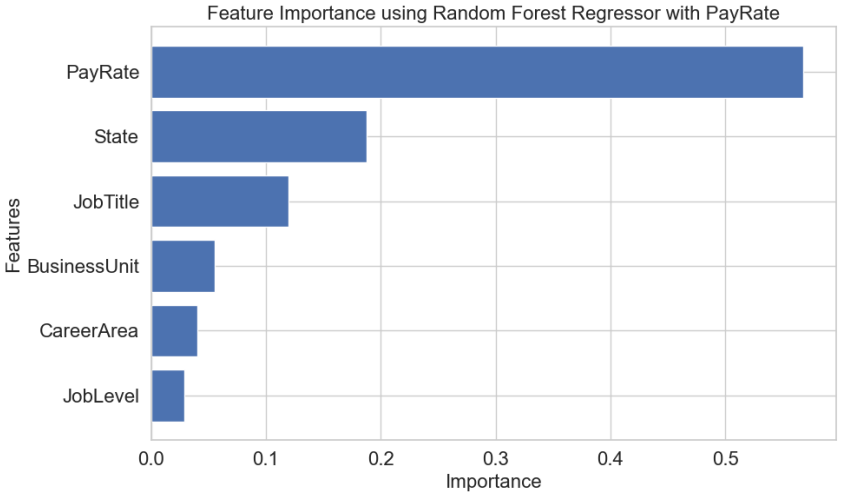
Random Forest	`max_depth=None,`
	`max_samples=1.0,`
	`min_samples_split=0.0001,`
	`n_estimators=500`

XGBoost	`learning_rate=0.3,`
	`max_depth=None,`
	`n_estimators=500,`
	`subsample=1.0`

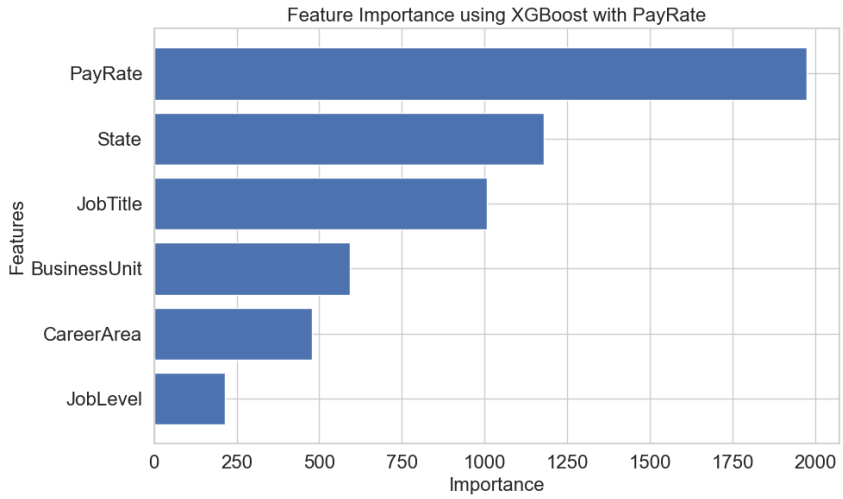
LightGBM	`learning_rate=0.1,`
	`n_estimators=500,`
	`num_leaves=127,`
	`subsample=0.6`

Table 8.6: Best Hyperparameters for Tree-based Models

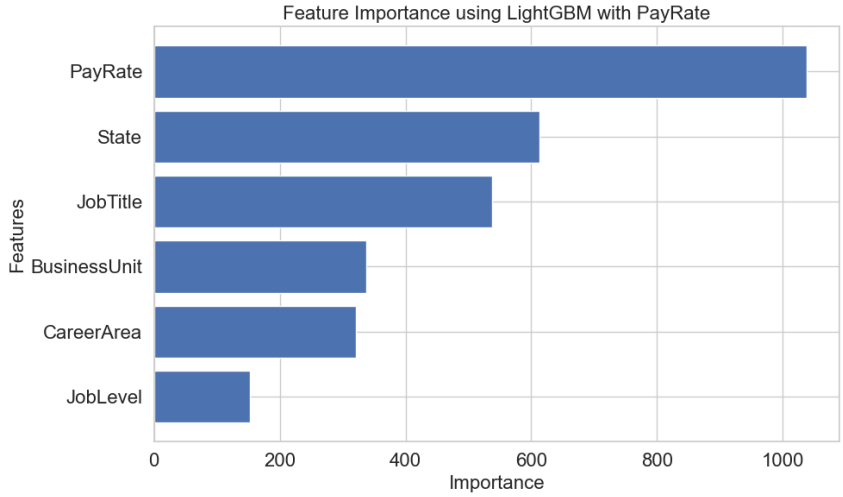
Through the previous work, the optimal hyperparameters of the three types



(a) Random Forest



(b) XGBoost



(c) LightGBM

Figure 8.1: Feature Importance using recommended hyperparameters

of models are obtained. Table 8.7 shows their performance on the test set.

Metric	Random Forest	XGBoost	LightGBM
RMSE	99.092	100.096	99.959
MAE	60.553	61.551	61.458
R^2	0.214	0.198	0.201
MSE	9819.261	10019.289	9991.901

Table 8.7: Performance Metrics for tree-based Models with the best hyperparameters

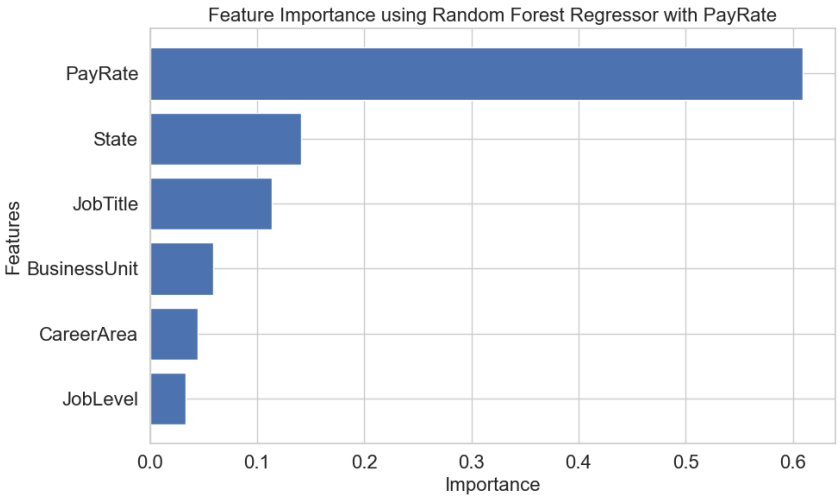
After tuning with hyperparameters, random forest, XGBoost, and LightGBM all achieved better performance.

We will find that XGBoost and LightGBM have a large performance improvement after hyperparameter tuning, which suggests that the performance of these two models on the work duration prediction problem is sensitive to the hyperparameters, whereas Random Forest is relatively insensitive.

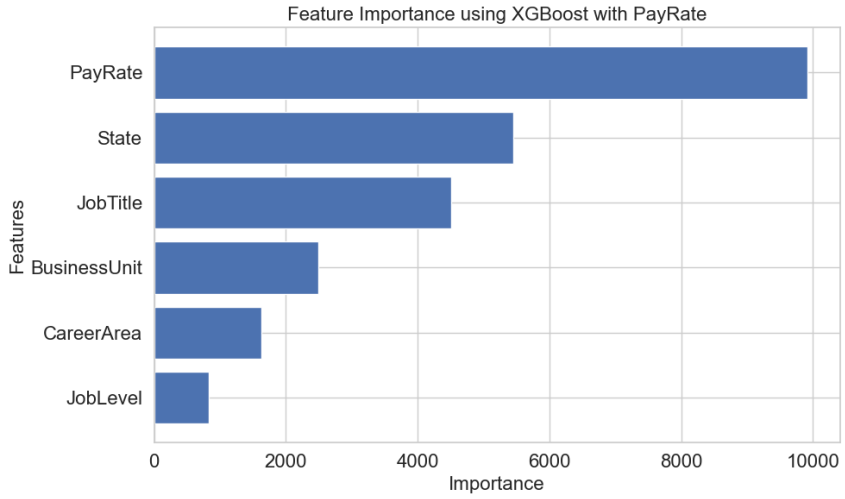
The feature importance of the model using the best hyperparameters of random forest, XGBoost, and LightGBM is shown in Figure 8.2.

It is worth noting that LightGBM achieves better performance, only lagging behind the Random Forest, but reduces its dependence on payrate (but payrate is still the most important) which leads us to believe that it learns a different prediction model than random forest.

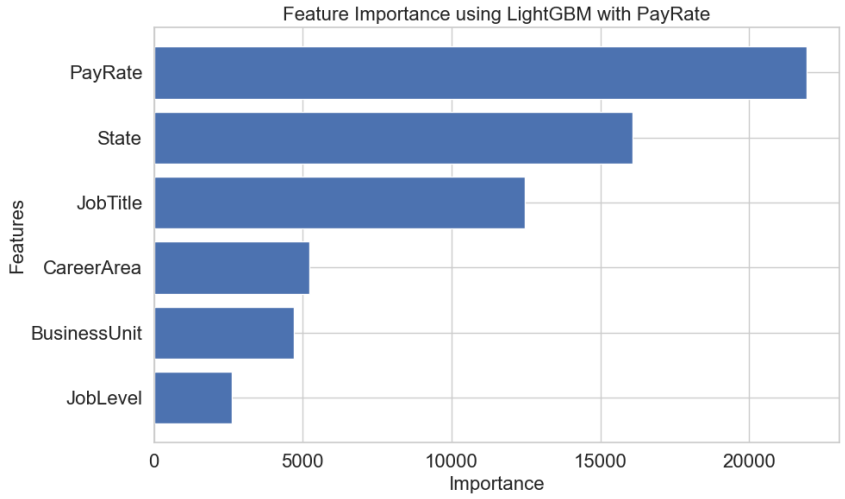
However, random forest still achieves the best performance.



(a) Random Forest



(b) XGBoost



(c) LightGBM

Figure 8.2: Feature Importance using best hyperparameters

9. Group Analysis

In order to give more intuitive explanations about the prediction effect of the pay-rate-based predictor we constructed, as well as an in-depth analysis of the prediction error, the next step is to group the duration of work, the business unit, the state, and the level of the job, and observe the performance of the prediction error of the best model with optimal hyperparameters, Random Forest, in the different groupings.

The results of the subgroup analyses help us to understand whether there are differences in the relationship between the effect of compensation on work duration in the different groupings.

9.1 Group by Work Duration

The AE (Absolute Error) percentage was categorized into three levels: "Low", "Medium", and "High", where "Low" corresponds to AE percentage less than 50%, "Medium" corresponds to AE percentage between 50% and 100%, and "High" corresponds to AE percentage exceeding 100%. Then, the dependent variable WorkDuration was divided into categories of one month, one quarter, half a year, and one year, and the sample sizes under different AE levels were counted.

After plotting the graph, it was observed that when WorkDuration was within one month, there were a considerable number of samples categorized under the "High" AE percentage level. This suggests that within a short work duration, the samples are relatively difficult to capture patterns, leading to larger prediction biases in the model, possibly influenced by various other factors. Conversely, for longer work durations, the model's predictive performance is better, especially in the 30-120 day interval, where

the majority of AE percentages are categorized as "Low", and even in durations exceeding 120 days, there were hardly any occurrences of the "High" AE level.

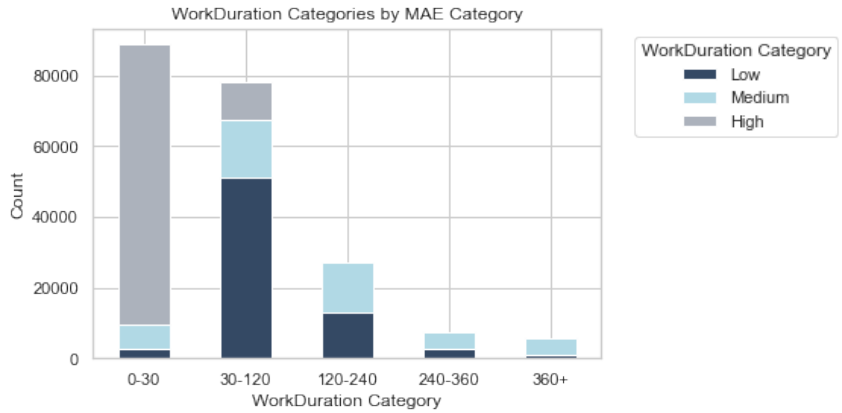


Figure 9.1: WorkDuration Categories by AE Category, the unit of the x-axis is days.

The absolute value prediction error is actually not favourable for evaluating the model as good or bad because of the large difference in the base value of work duration. Again, with a 60-day forecast error, when the work duration base is 100 days, this is a mediocre prediction, but if the base is 1000 days, this is an accurate prediction. Therefore, we constructed the "AE percentage", which also gives us another perspective on how good the model is.

In this section, we can see that if a job tends to be short-term, then its compensation is a relatively weak predictor of work duration. It is very likely that other more important factors determine the work duration of short-term workers. It is also possible that, due to the nature of some jobs, their work duration is necessarily short, independent of factors such as compensation. It is difficult to explore this further due to the limited dimensions of the data, but it is an interesting direction to consider!

In addition to this, the predictive effect of compensation on workduration is very strong in medium-length jobs (30 days-240 days), and most of the AE percentages of the predictive models we constructed fall in the Low interval. And medium-length jobs make up the main part of all jobs. This suggests

that our model largely accounts for the predictive role of compensation on workduration.

Using the dependent variable as a basis for grouping may lead to some confusion, but it provides a meaningful analytical perspective. Some jobs inherently have long, medium, or short durations. If historically the average length of a particular job has been medium, then our prediction is more likely to be accurate. Conversely, this analytical perspective suggests that if our prediction system gives a medium-length prediction, the probability of that prediction being accurate should also be relatively high.

9.2 Group by Business Unit

Similarly, categorizing the AE percentage into different levels and comparing the prediction errors across different business units revealed that the majority of the data was concentrated in RNALP, with small proportions in RISUS, SPHUS, RTDUS, and RPUUS. Notably, the predictive model demonstrated consistent performance across all business units, indicating uniform effectiveness in predicting work durations irrespective of the specific business unit.

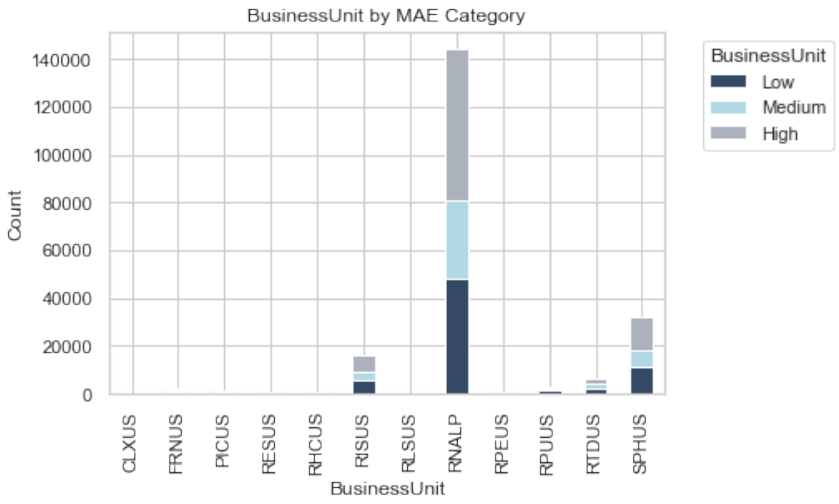


Figure 9.2: BusinessUnit by AE Category

For each business unit, random forest regression was applied separately to obtain these performance metrics, shown in Table 9.1. When considering

the R-squared values for the five business units with larger samples, RISUS exhibited the highest R-squared value, indicating that within this business unit, the salary level has the strongest explanatory power for work duration, followed by RNALP, SPHUS and RPUUS are slightly inferior. It is worth noting that the R-squared value for RTDUS is negative, suggesting that we did not identify an effective predictive pattern. In this subgroup, the salary level does not predict work duration effectively.

Business_Unit	RMSE	MAE	R^2	MSE
RISUS	91.978222	55.864465	0.191376	8459.993243
RNALP	91.878682	56.588544	0.174611	8441.69223
FRNUS	127.273024	71.27605	0.230096	16198.422641
PICUS	265.319192	169.628384	0.14638	70394.273626
RESUS	192.277865	118.943984	0.236865	36970.777379
SPHUS	82.756359	49.879276	0.123419	6848.614942
RPUUS	131.811724	88.059151	0.075421	17374.330575
RHCUS	114.67051	81.378924	0.133173	13149.325899
RTDUS	229.212219	158.986162	-0.132093	52538.241499
RLSUS	194.50751	144.916338	-0.004074	37833.171404
RPEUS	165.853914	120.094063	-0.093481	27507.520878
CLXUS	82.397578	70.597728	-0.045755	6789.360928
RPOUS	91.0	91.0	NaN	8281.0

Table 9.1: Metrics for Random Forest across different business units

9.3 Group by State

Comparing the prediction errors across different states revealed that the majority of data was distributed in CA, FL, GA, NC, TN and TX, with small proportions in other states. The predictive model exhibited similar performance across all states, with comparable proportions of samples categorized under different AE levels.

For each state, random forest regression was applied separately to obtain these performance metrics, shown in Table 9.2. In the examination of R-squared values where sample sizes are larger, West Virginia (WV) exhibits the best explanatory power of salary levels on work duration, followed by

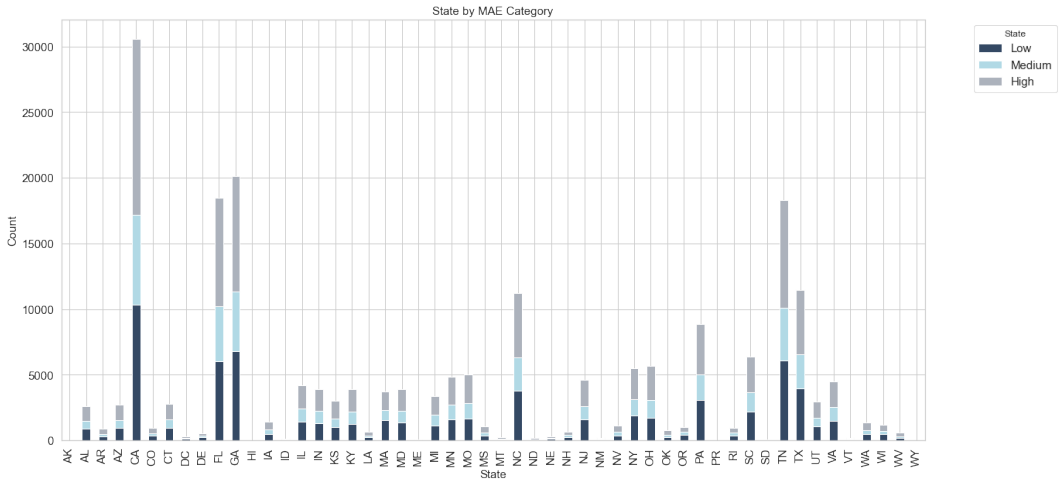


Figure 9.3: State by AE Category

Pennsylvania (PA), Kentucky (KY), and Maryland (MD), which all demonstrate good predictive capabilities. In the District of Columbia (DC), Colorado (CO), Kansas (KS), Virginia (VA), Connecticut (CT), and Louisiana (LA), the R-squared values do not exceed 10%, indicating relatively average explanatory power. Meanwhile, negative R-squared values are observed in New Mexico (NM), South Dakota (SD), and Hawaii (HI), which suggests that the forecasts are underperforming and that there are different patterns of determining workduration in these states.

9.4 Group by Job level

Comparing the prediction errors across different job levels revealed that the majority of the data was concentrated at level 1.0, entry level, with small proportions in other levels. Notably, the predictive model demonstrated similar performance across all job levels, with comparable proportions of samples categorized under different AE levels.

For each job level, random forest regression was applied separately to obtain these performance metrics, shown in Table 9.3. Upon analyzing the R-squared values for different job levels, it is observed that the explanatory power of salary on work duration shows a U-shape pattern. The highest job level demonstrates the strongest explanatory capacity, followed by the

State_	RMSE	MAE	R^2	MSE
IL	120.966316	69.516496	0.203405	14632.849635
DC	209.626991	123.747991	0.002744	43943.475414
GA	87.971382	54.940901	0.141381	7738.964081
NJ	136.237458	84.859485	0.151364	18560.644895
AL	103.7202	56.270235	0.22958	10757.879866
NY	146.027037	86.563162	0.18683	21323.895552
AZ	110.934124	72.552459	0.109292	12306.379858
CA	97.076693	59.19385	0.154655	9423.884227
PA	104.669527	61.711826	0.251125	10955.70985
MD	116.925239	68.275739	0.243953	13671.511575
FL	103.639702	61.167968	0.169265	10741.187876
MO	80.082241	47.491545	0.197968	6413.165248
MI	86.47188	54.230127	0.135886	7477.385973
KY	116.317414	64.114934	0.244092	13529.740696
IA	98.032688	58.339251	0.106457	9610.407912
TX	108.426021	65.903689	0.12626	11756.202133
SC	84.220973	52.827651	0.138309	7093.172325
MT	124.616616	75.525302	0.153826	15529.301087
MN	100.618441	59.506005	0.140436	10124.070613
WI	116.6678	70.249131	0.202653	13611.375501
CO	126.323711	71.828743	0.076017	15957.679981
TN	77.701495	49.407722	0.103938	6037.522268
KS	59.1346	41.858379	0.056473	3496.900941
MA	156.00195	93.791132	0.185728	24336.608476
VA	111.008304	65.131321	0.057624	12322.843494
NC	101.88314	59.643574	0.171983	10380.174182
WA	126.193763	73.396032	0.209759	15924.86592
CT	107.14838	67.198688	0.073174	11480.775407
OH	91.651189	53.62176	0.167672	8399.940414
DE	151.837093	105.336593	0.200473	23054.50267
OR	114.163148	77.612824	0.187938	13033.224263
NE	67.334631	41.860354	0.174465	4533.952503
NH	106.044413	68.094367	0.116453	11245.417571
RI	117.002003	76.274474	0.154469	13689.468648
OK	115.179026	59.063026	0.159371	13266.208107
MS	81.820397	57.761042	0.123843	6694.577325
IN	66.880407	43.742473	0.144776	4472.988903
ME	260.215895	164.750028	0.113743	67712.312005
AR	79.509005	48.815311	0.133747	6321.681804
NV	93.466162	55.498811	0.146293	8735.923356
LA	133.369945	67.922127	0.035616	17787.54228
UT	88.366233	54.310571	0.139449	7808.591222
NM	97.459537	65.350155	-0.035759	9498.361351
WY	229.079985	177.038739	0.200832	52477.639497
AK	134.306276	79.333179	0.48383	18038.175796
SD	118.31695	77.884317	-0.5476	13998.900737
WV	79.15729	46.755973	0.336528	6265.876609
HI	272.344402	228.414733	-0.888949	74171.473507
ID	101.874258	61.291571	0.031298	10378.364481
ND	90.333678	58.555322	0.11339	8160.173403
VT	129.413463	73.523255	0.16936	16747.844353
NaN	0.182	0.182	NaN	0.033124
PR	0.0	0.0	NaN	0.0

Table 9.2: Metrics for Random Forest across different states

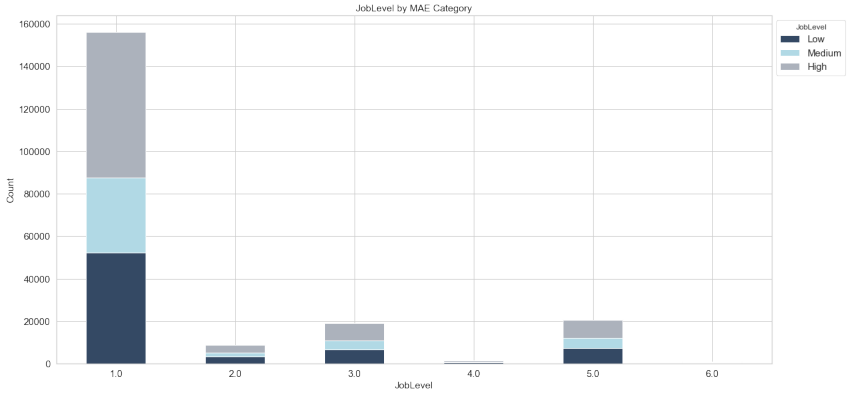


Figure 9.4: JobLevel by AE Category

lowest job level which also exhibits substantial explanatory strength. However, when the job level is at a medium degree, specifically at level 4, the explanatory power is the weakest.

JobLevel	RMSE	MAE	R^2	MSE
3.0	118.517257	70.102751	0.143725	14046.340187
1.0	85.364512	53.490232	0.182607	7287.09996
2.0	132.159162	78.724433	0.148127	17466.043985
6.0	173.960789	106.839449	0.210661	30262.355995
4.0	179.619761	117.369511	0.022855	32263.258391
5.0	147.710397	87.658547	0.151527	21818.361334

Table 9.3: Metrics for Random Forest across different job levels

The group analysis of the payrate-based predictor provides insights into the prediction performance and error across various dimensions, such as work duration, business unit, state, and job level. Grouping by work duration revealed that shorter work durations tend to have higher prediction errors, indicating that compensation is a weaker predictor for these jobs, potentially due to other influential factors. In contrast, medium-length jobs showed stronger predictive performance, suggesting the model effectively captures the compensation's impact on work duration for these jobs. Grouping by business unit highlighted consistent predictive performance across most units, with the highest explanatory power observed in RISUS and RNALP units. State-level analysis showed that while some states like West Virginia and Pennsylvania exhibited strong predictive capabilities, others had weaker or even negative explanatory power, indicating varying effectiveness of the model across different states. Lastly, grouping by job level demonstrated a U-shaped pattern, where the highest and lowest job levels had strong explanatory power, whereas medium job levels, particularly level 4, showed weaker prediction accuracy. Overall, these subgroup analyses help identify the conditions under which the payrate-based predictor performs well and where it might need further refinement.

10. Discussion

10.1 Quantification of the impact of compensation on work duration

To what extent can compensation determine or predict work duration? This directly addresses our first research question (RQ1: To what extent does compensation predict work duration?). Until now, we have used a variety of statistical tools to verify the notable impact of compensation on work duration. The quantitative results provide strong evidence of its implications, including:

- **Initial explorations of linear regression:** Our analysis began with linear regression, confirming a positive correlation from multiple perspectives, and highlighted that compensation impacts different job categories variably.
- **Feature importance of non-linear models:** Construct different tree models to discover more complex non-linear relationships and calculate feature importance, demonstrating that pay rate is a pivotal feature in predicting work duration.
- **Paired model test:** Different tree models with and without the payrate variable were constructed with maximum likelihood ratio tests and paired t-tests, and the results showed that the model with the payrate variable performed better, reinforcing the predictive importance of compensation.
- **Good predictive performance with payrate as the main feature:** In fact, except for records with work durations of less than one month, the best Random Forest model relies on payrate as the main feature and predicts most of the records well, falling in the "Low AE" range.

In contrast, for work records less than one month old, which form the main component of the dataset, the predictive models built have basic predictive ability but are not as remarkable.

- **Differences between gig work and traditional work:** For work records shorter than one month, the primary dataset component, the constructed predictive models exhibit fundamental but unremarkable predictive abilities. This underscores the need for developing innovative compensation strategies within the gig economy. The dynamics between payrate and employee retention periods may differ between gig jobs and more traditional roles, presenting unique challenges for prediction. This situation necessitates the collection of more relevant data fields to identify and leverage predictive patterns more effectively.

10.2 Performance of predicted effects on different groups

To what degree and in what way does compensation impact the length of employment across different work duration, business units, states, and job levels? This addresses our second research question (RQ2: Does the predictability of compensation for work duration vary across different levels within the labor market?). By categorizing error percentages and observing the predictive outcomes in different groups along with the explanatory power reflected by the R-squared values, the following findings were obtained:

- **Group by Work Duration:** The predictive performance of our model exhibits large differences across various work duration groupings. The prediction model we constructed for medium-length work records performs well when payrate is used for work duration, which is the main part of all work. However, in short-term (less than 30 days) and long-term (more than 240 days) work records, this predictive relationship becomes much weaker, either due to other determinants or the inherent characteristics of certain jobs. A more field-rich dataset would also

be needed to investigate the rationale.

- **Group by Business Units:** The distribution of error levels across various business units is relatively similar, but some business units have better predictive performance and some have very weak predictive relationships. Salary within RISUS has the strongest explanatory power for work duration, whereas in RTDUS, our model's predictions are not effective. Different business units correspond to different job characteristics and hiring patterns, and further exploration of the reasons for this can help promote better hiring strategies.
- **Group by States:** The distribution of error levels is relatively similar across different states. However, some states have better predictive performance and some have very weak predictive relationships. Salary has the strongest prediction power for work duration in West Virginia (WV), Pennsylvania (PA), Kentucky (KY), and Maryland (MD), also performing well. However, the prediction power is very poor in the District of Columbia (DC), Louisiana (LA), and Idaho (ID). The consistency of model performance across states with notable exceptions (e.g., West Virginia vs. the District of Columbia) suggests regional economic conditions, labor laws, or labor market differences among other states might influence the compensation-work duration relationship.
- **Group by Job Levels:** The distribution of error levels is relatively similar across different job levels. However, it is noteworthy that the prediction power of salary on work duration exhibits a U-shaped pattern. The R-squared values are higher at the highest and lowest job levels, whereas at medium job levels, the R-squared values are very small, indicating poor prediction power. This can guide us to develop different compensation strategies for different job levels.

11. Conclusion

This study explores the relationship between compensation and work duration across various job categories, particularly focusing on linear regression and tree-based models for predictive analysis. It finds that, in most jobs, compensation has a statistically positive effect on work duration. Higher salaries are linked to longer employment periods, especially in lower-level, blue-collar positions. The findings suggest that well-structured compensation packages can enhance employee retention and satisfaction.

This research offers useful insights for human resources professionals and organizational leaders, emphasizing the importance of tailoring compensation strategies to not only attract talent but also to foster long-term employment relationships. In an era where talent retention is as crucial as talent acquisition, such insights can be used to develop fair compensation packages that are aligned with both organizational goals and employee expectations.

Although this study provides a basis for understanding compensation's influence on work duration, it acknowledges certain limitations. Future studies should incorporate data from a wider range of sectors, including white-collar and permanent positions, to enhance the generalizability of the findings.

Moreover, the exploration of additional variables that may influence the relationship between compensation and work duration, such as work-life balance, job satisfaction, and the role of benefits beyond base salary, is promising for future research. Integrating qualitative methods to complement the quantitative analysis could also uncover deeper insights into the motivations behind employee tenure decisions.

A. Appendix

A.1 Randstad USA dataset

This appendix details the dataset used in our study. The data was sourced from Randstad USA, a prominent workforce solutions provider, and encompasses extensive historical assignment data. This dataset was accessed through Google BigQuery and includes millions of entries spread across a wide range of variables. These variables capture detailed information about job assignments, including pay rates, job titles, job levels, geographic locations, industry sectors, and more. The purpose of this detailed schema is to ensure clarity and provide a comprehensive understanding of the dataset used, which is crucial for the analysis performed in the study.

Data Fields

- **RNA_REPORT_DATE**:

  Description: The date and time when the data was entered into the database.

  Data Type: DateTime

  Format: YYYY-MM-DD HH:MM:SS UTC

- **BUSINESS_UNIT**:

  Description: The business unit responsible for the management and operations of the assignment.

  Data Type: String

- **ASSIGNMENT_ID**:

  Description: A unique identifier for the specific job assignment.

  Data Type: String

- **PayRate**:

  Description: The hourly wage rate for the assignment.

  Data Type: Float

- **JOB_TITLE**:

  Description: The official job title of the assignment.

  Data Type: String

- **CITY**:

  Description: The city where the job assignment is located.

  Data Type: String

- **STATE**:

  Description: The U.S. state where the job is located.

  Data Type: String

- **CUST_NAME**:

  Description: The name of the customer or company where the assignment is located.

  Data Type: String

- **clean_job_title**:

  Description: A simplified or formatted version of the job title for analytical consistency.

  Data Type: String

- **normalized_job_title**:

  Description: A standardized version of the job title used for categorization.

  Data Type: String

- **score**:

  Description: A confidence score representing the accuracy of the job title normalization.

Data Type: Float

Range: 0 to 1

- **job_level_us**:

  Description: Indicates the job level, related to experience and responsibilities.

  Data Type: Integer

  Range: 1 (entry-level) to 6 (senior-level)

- **careerarea_name_us**:

  Description: The career area or industry category of the job.

  Data Type: String

A.2 Hyperparameter Grid Search for Model Tuning

This appendix provides an overview of the hyperparameters explored for three different models: Random Forest Regressor, XGBoost Regressor, and LightGBM Regressor. For each model, we employed GridSearchCV to identify the best combination of hyperparameters to minimize the mean squared error.

RandomForestRegressor

The `RandomForestRegressor` model was tuned using the following hyperparameters:

- `n_estimators`: Number of trees in the forest.
  - Values: [100, 200, 500]
- `max_depth`: Maximum depth of the tree.
  - Values: [None, 10, 20, 30]
- `min_samples_split`: Minimum number of samples required to split an internal node.
  - Values: [0.001, 0.01, 0.05, 0.1]
- `max_features`: Number of features to consider when looking for the best split.
  - Values: ['sqrt', 'log2', 0.5]
- `max_samples`: Fraction of samples to be used for fitting each individual base learner.
  - Values: [0.6, 0.8, 1.0]

XGBoostRegressor

The `XGBoostRegressor` model was tuned using the following hyperparameters:

- `n_estimators`: Number of boosting rounds.
  - Values: [100, 200, 500]
- `max_depth`: Maximum depth of a tree.
  - Values: [None, 10, 20, 30]
- `learning_rate`: Step size shrinkage used to prevent overfitting.
  - Values: [0.01, 0.05, 0.1]
- `subsample`: Subsample ratio of the training instance.
  - Values: [0.5, 0.7, 1.0]

LGBMRegressor

The `LGBMRegressor` model was tuned using the following hyperparameters:

- `n_estimators`: Number of boosting rounds.
  - Values: [100, 200, 500]
- `num_leaves`: Maximum tree leaves for base learners.
  - Values: [31, 62, 127]
- `learning_rate`: Step size shrinkage used to prevent overfitting.
  - Values: [0.01, 0.05, 0.1]
- `subsample`: Subsample ratio of the training instance.
  - Values: [0.6, 0.8, 1.0]

Bibliography

[1] A. Frimayasa, "Effect of compensation, career development and work environment on employee retention (study on employees of pt telkom witel tangerang bsd)," *Journal of Research in Business, Economics, and Education*, vol. 3, no. 1, pp. 1715–1730, 2021.

[2] N. Gupta and J. D. Shaw, "Employee compensation: The neglected area of hrm research," *Human resource management review*, vol. 24, no. 1, pp. 1–4, 2014.

[3] S. Steinmetz, D. H. d. Vries, and K. G. Tijdens, "Should i stay or should i go? the impact of working time and wages on retention in the health workforce," *Human resources for health*, vol. 12, pp. 1–12, 2014.

[4] U. N. Urme, "The impact of talent management strategies on employee retention," *International Journal of Science and Business*, vol. 28, no. 1, pp. 127–146, 2023.

[5] I. a. Setiawan, S. Suprihanto, A. Nugraha, and J. Hutahaean, "Hr analytics: Employee attrition analysis using logistic regression," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 830, 2020, p. 032 001.

[6] L. Yu, X. Zhao, J. Huang, H. Hu, and B. Liu, "Research on machine learning with algorithms and development," *Journal of Theory and Practice of Engineering Science*, vol. 3, no. 12, pp. 7–14, 2023.

[7] F. Eichinger and M. Mayer, "Predicting salaries with random-forest regression," in *Machine Learning and Data Analytics for Solving Business Problems: Methods, Applications, and Case Studies*, Springer, 2022, pp. 1–21.

[8] P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *algorithms*, vol. 4, no. 5, p. C5, 2016.

[9] K. M. Kuhn, J. Meijerink, and A. Keegan, "Human resource management and the gig economy: Challenges and opportunities at the intersection between organizational hr decision-makers and digital labor platforms," *Research in personnel and human resources management*, vol. 39, pp. 1–46, 2021.

[10] R. Malik, A. Visvizi, and M. Skrzek-Lubasińska, "The gig economy: Current issues, the debate, and the new avenues of research," *Sustainability*, vol. 13, no. 9, p. 5023, 2021.

[11] *What exactly is the gig economy | Randstad*, https://www.randstad.com/workforce-insights/future-of-work/what-exactly-gig-economy/. (visited on 06/08/2024).

[12] *Temp & Staffing Agency for Job Seekers & Employers | Randstad USA*, https://www.randstadusa.com/. (visited on 06/07/2024).

[13] I. Lipovac and M. B. Babac, "Developing a data pipeline solution for big data processing," *International Journal of Data Mining, Modelling and Management*, 2024.

[14] *Burning Glass Occupation Taxonomy processes - OrphanItems - Confluence*, https://economicmodeling.atlassian.net/wiki/spaces/SPC/pages/2601780294/ (visited on 06/09/2024).

[15] scikit-learn developers, *sklearn.preprocessing.LabelEncoder*, `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html`, Accessed: 2023-09-10, 2023.

[16] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[17] M. R. Segal, "Machine learning benchmarks and random forest regression," 2004.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[19] G. Ke, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[20] R. Protassov, D. A. Van Dyk, A. Connors, V. L. Kashyap, and A. Siemiginowska, "Statistics, handle with care: Detecting multiple model components with the likelihood ratio test," *The Astrophysical Journal*, vol. 571, no. 1, p. 545, 2002.

[21] H. Hsu and P. A. Lachenbruch, "Paired t test," *Wiley StatsRef: statistics reference online*, 2014.

[22] M. Padhma, "A comprehensive introduction to evaluating regression models," *Data Science Blogathon updated On October 31st*, 2023.

[23] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: journal of the Econometric Society*, pp. 307–333, 1989.

[24] Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[26] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, "Experimenting xgboost algorithm for prediction and classification of different datasets," *International Journal of Control Theory and Applications*, vol. 9, no. 40, pp. 651–662, 2016.

[27] C. XGBoost, S. LightGBM, and B. Quinto, *Next-generation machine learning with spark*, 2020.

[28] B. D. Ripley, *Pattern recognition and neural networks*. Cambridge university press, 2007.

[29] M. Feurer and F. Hutter, "Hyperparameter optimization," *Automated machine learning: Methods, systems, challenges*, pp. 3–33, 2019.

[30] M. Kuhn, "Caret: Classification and regression training," *Astrophysics Source Code Library*, ascl–1505, 2015.

[31] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, 2019.

[32] G. Ranjan, A. K. Verma, and S. Radhika, "K-nearest neighbors and grid search cv based real time fault monitoring system for industries," in *2019 IEEE 5th international conference for convergence in technology (I2CT)*, IEEE, 2019, pp. 1–5.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[34] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of cheminformatics*, vol. 6, pp. 1–15, 2014.

[35] S. Bleeker, H. Moll, E. a. Steyerberg, *et al.*, "External validation is necessary in prediction research:: A clinical example," *Journal of clinical epidemiology*, vol. 56, no. 9, pp. 826–832, 2003.

[36] E. W. Steyerberg, S. E. Bleeker, H. A. Moll, D. E. Grobbee, and K. G. Moons, "Internal and external validation of predictive models: A simulation study of bias and precision in small samples," *Journal of clinical epidemiology*, vol. 56, no. 5, pp. 441–447, 2003.

[37] Y. Chen, I. Moustaki, and H. Zhang, "A note on likelihood ratio tests for models with latent variables," *psychometrika*, vol. 85, no. 4, pp. 996–1012, 2020.

[38] H. A. David and J. L. Gunnink, "The paired t test under artificial pairing," *The American Statistician*, vol. 51, no. 1, pp. 9–12, 1997.

[39] P. R. Killeen, "Replication statistics," *Best practices in quantitative methods*, pp. 103–124, 2007.

[40] E. Duarte and J. Wainer, "Empirical comparison of cross-validation and internal metrics for tuning svm hyperparameters," *Pattern Recognition Letters*, vol. 88, pp. 6–11, 2017.