**Project Milestone 5 – Final**

**Introduction**

**Problem Statement:**

The project seeks to address a critical challenge faced by our organization - forecasting salary projections related to total human capital. To achieve this, I utilized a comprehensive payroll salary dataset spanning from January 2021 to May 2023. The ultimate objective was to leverage the power of data-driven insights to enhance the accuracy and effectiveness of forecasting human capital salary expense. As our organization incurs substantial expenses in salaries and related costs, which amount to over $40M on a revenue book of $550M+, accurate salary predictions are significant. By obtaining reliable projections of future payroll expenses, we can make well-informed financial decisions, optimize resource allocation, and strategically plan.

**Importance of the Problem:**

This endeavor holds significant implications for our company's success. Effective salary forecasting empowers us to engage in efficient budgeting and resource allocation, ensuring we remain competitive in the market. Moreover, it enables us to adopt proactive workforce planning strategies, fostering compensation equity and career development paths for our valued employees. By leveraging historical payroll data and applying advanced predictive modeling techniques, I aimed to build a robust model capable of making accurate and precise salary projections. The success of this project will empower our organization with data-driven insights, enabling us to drive optimal business outcomes, attract and retain top talent, and align our compensation practices with industry standards and market conditions. The project's successful execution will contribute to the long-term growth and sustainability of our family-owned business, solidifying our position as a leading force in the market. Through the implementation of advanced data analytics, we aim to make informed decisions that drive profitability, nurture a thriving work environment, and secure our organization's future success.

**Target Audience:**

The project's insights and findings will be of importance to key stakeholders within our organization, including those responsible for financial planning, human resources, and executive decision-making. By having access to accurate salary projections and comprehensive human capital insights, these stakeholders can make well-informed and data-driven decisions based upon our actual expense experience.

Additionally, external parties, such as investors, analysts, and consultants, will find value in the project's outcomes. The ability to assess the organization's financial health and strategic direction based on reliable salary forecasting and human capital analysis will enable them to make informed investment decisions, evaluate our long-term sustainability, and offer valuable recommendations for further growth and success.

**Data Source:**

The primary data source for this project is the Human Resources Information System (HRIS) platform provided by ADP (Automatic Data Processing). The dataset comprises comprehensive payroll salary data spanning from January 2021 to May 2023, capturing essential information such as payroll dates, amounts, employee attributes, and job characteristics. The HRIS platform ensures the data's accuracy, reliability, and compliance with data privacy regulations, making it a trustworthy and robust source for my analysis. By leveraging this extensive and well-organized dataset, I can extract meaningful insights and develop predictive models that will enhance our understanding of salary projections.

**Usefulness of the Data:**

The data obtained from ADP holds significance for this project, as it forms the backbone of my ability to build powerful predictive models. Its usefulness lies in several key aspects:

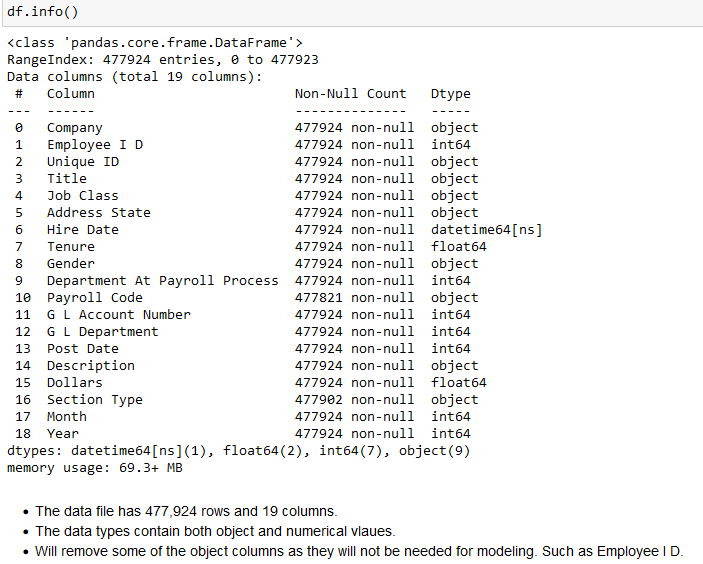
1. Accurate Salary Forecasting: The historical payroll transactions contained in the dataset provide a rich and diverse set of data points that enables me to create robust predictive models. By leveraging this data, I can develop algorithms that accurately forecast future salaries for employees, departments, and the organization.
2. Informed Financial Planning: The ability to predict future salary expenses with precision empowers our financial planning efforts. With accurate salary projections at our disposal, we can optimize budget allocation, plan for operational costs, and ensure financial stability in the short and long term.
3. Data-Driven Decision-Making: In a competitive business landscape, data-driven decision-making is crucial. The insights derived from the data analysis allow us to make informed choices regarding compensation, talent management strategies, and workforce planning. It ensures that our decisions align with organizational goals and benchmark to market conditions.
4. Talent Management and Retention: Salary forecasting enables us to develop compensation equity measures, ensuring fair and competitive wages for our employees. By making data-backed decisions, we can attract and retain top talent, fostering a motivated and engaged workforce.
5. Strategic Business Decisions: The ability to predict payroll expenses accurately has broader implications for the overall business strategy. It allows us to assess the financial impact of hiring new employees, adjusting compensation structures, or implementing cost-saving measures.
6. Continuous Improvement: By analyzing historical payroll data, I can identify trends, patterns, and outliers that shed light on potential areas for improvement. These insights can be used to refine our business processes and enhance operational efficiency.
7. Compliance and Accountability: The dataset from ADP ensures data compliance with applicable privacy and security regulations. It enables us to maintain transparency, accountability, and adherence to legal requirements in all aspects of compensation and payroll management.

The data from ADP is indispensable for developing accurate predictive models, supporting financial planning, driving data-driven decision-making, and fostering a fair and competitive compensation strategy. By harnessing the potential of this dataset, we position our organization for sustainable growth, strategic excellence, and competitive advantage in the market.

**Methods/Results**

**Exploration of Data:**

During the initial phase of data exploration, I conducted a thorough examination of the dataset to gain insights into its structure and quality. Missing values were identified in some uncoded payroll entries, which were subsequently removed from the dataset due to their minimal impact on the overall data integrity. The remaining data proved to be well-organized, with each row representing a distinct employee, creating an ideal foundation for analysis.



**Data Dictionary**

| **Column Name** | **Description** |
| --- | --- |
| Company | The name or identifier of the company that the employee belongs to. |
| Employee ID | A unique identifier for each employee in the company. |
| Unique ID | Another unique identifier, used for specific purposes within the dataset. |
| Title | The job title or position of the employee within the company. |
| Job Class | The job classification or category of the employee's position. |
| Address State | The state where the employee's address is located. |
| Hire Date | The date when the employee was hired by the company. |
| Tenure | The length of time (in years) the employee has been working in the company. |
| Gender | The gender of the employee. |
| Department At Payroll Process | The department to which the employee belongs at the time of payroll processing. |
| Payroll Code | A code related to the employee's payroll transactions. |
| G L Account Number | The account number used in the company's General Ledger for financial payroll transactions. |
| G L Department | The department in the company associated with financial transactions. |
| Post Date | The date when a transaction or entry was posted. |
| Description | A description of the check date. |
| Dollars | The amount of money involved in the transaction. |
| Section Type | The type or category of the payroll code. |
| Month | The month in which a particular event or transaction occurred. |
| Year | The year in which a particular event or transaction occurred. |

**Data Preparation:**

Data preparation played a pivotal role in ensuring the dataset's suitability for predictive modeling. I meticulously cleaned the dataset, removing redundant columns that contributed little to the model's predictive power. Essential categorical variables, such as Title, Job Class, Address State, and Gender, were retained, as they proved to be influential factors in salary forecasting. By optimizing the dataset, I streamlined the modeling process, leading to more accurate predictions.

A screenshot of a computer

Description automatically generated

The code below demonstrates the accounts associated with the GL Numbers I used to formulate the “base” earnings.

A screenshot of a computer

Description automatically generated

**Visualizations:**

Various exploratory visualizations were employed to gain deeper insights into the data. Box plots and bar plots were utilized to identify potential salary distribution disparities and discern the compensation hierarchy among different job classes. These visualizations revealed valuable patterns and trends, aiding in the formulation of data-driven decisions. Additionally, feature engineering techniques and a correlation matrix were applied to uncover potential correlations and dependencies within the data, enhancing our understanding of the factors influencing salaries.

A graph of blue squares

Description automatically generated

**Modeling:**

In pursuit of the most accurate predictive model, I explored three different algorithms: Linear Regression, Random Forests, and Gradient Boosting Models (specifically XGBoost). Each model was evaluated based on its performance, accuracy, and ability to handle complex patterns within the data. After comprehensive evaluation, XGBoost emerged as the optimal choice, surpassing the other models in terms of accuracy and performance. Its ability to handle intricate relationships within the data and effectively prevent overfitting made it the most suitable algorithm for our salary forecasting task.

**Metrics:**

To evaluate the models, I employed a range of well-established metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2). These metrics allowed us to assess each model's predictive capabilities and gauge their performance. XGBoost demonstrated superior performance with the lowest MAE and the highest R2 score, indicating its effectiveness in accurately predicting future salaries. The metrics validated our choice of the XGBoost model as the most reliable and robust approach for this project.

A screenshot of a computer error

Description automatically generated

**Conclusion**

**Key Learnings:**

The project has been a journey, yielding valuable insights that have significant implications for our organization's success and long-term growth. Through rigorous analysis and data-driven methodologies, several key learnings have emerged:

1. Accurate Salary Forecasting: The project reaffirmed the criticality of accurate salary forecasting in achieving optimal financial planning and resource allocation. By leveraging historical payroll data and sophisticated predictive models, we now possess the ability to make precise salary projections, aiding in budgeting and financial decision-making. A prediction was made to be accurate within 2% of actuals for the month of June 2023 for the entire corporation.
2. Compensation Equity and Fairness: The project shed light on the importance of compensation equity within our organization. By analyzing salary distribution disparities and implementing data-backed decisions, we can ensure fair and competitive wages for all employees, fostering a positive work culture and enhancing employee satisfaction. A prediction was done for a test input where all things remained the same apart from changing the binary coding for male (M) to female (F) and the outcome was within 1% of equality.
3. Data-Driven Decision-Making: Embracing a data-driven approach has proven to be instrumental in shaping strategic decisions. The project demonstrated the power of data analytics in guiding executive decision-making, enhancing our ability to plan, strategize, and achieve our business objectives effectively.
4. Business Strategy Optimization: The project's findings have provided us with a deeper understanding of the factors influencing salary expenses. This knowledge can be leveraged to optimize our business strategy, identifying areas for cost-saving measures and areas for investment in talent development.
5. Continuous Improvement: The project emphasized the significance of continuous improvement in data analytics and modeling techniques. As we evolve, it is essential to regularly update and refine my models to align with changing business conditions and market dynamics.
6. Collaboration and Cross-Functional Insights: The project's continued success will be bolstered by cross-functional collaboration between finance, human resources, and data analytics teams.

Moving forward, these key learnings will serve as a compass, guiding us toward making data-driven, equitable, and strategic decisions.

**Recommendations:**

To propel the model's performance and ensure its continued relevance, a set of valuable recommendations have emerged from the project's insights and learnings. By incorporating these enhancements, I can further optimize our salary forecasting model and foster data-driven decision-making:

1. Assumed Annual Merit Increases: To improve the model's accuracy and realism, incorporating assumed annual merit increases is crucial. As salaries often undergo adjustments based on employee performance and market conditions, integrating these assumptions will allow us to better reflect real-world scenarios and anticipate future payroll expenses more accurately.
2. Accounting for Headcount Changes: The workforce is dynamic, and headcount fluctuations are commonplace in any organization. By considering the impact of headcount additions or deletions throughout the year, we can refine the model's predictions to align with changing personnel requirements and workforce planning strategies.
3. Exploring External Factors: Beyond internal data, external factors can significantly influence salary expenses. Exploring external variables such as economic indicators, market trends, and industry-specific factors can offer a more comprehensive understanding of salary projections. Integrating these external factors into the model can enhance its predictive power and strategic relevance.
4. Continuous Model Refinement: Continuously refining and updating the model is essential to maintain its accuracy and effectiveness. Regularly evaluating the model's performance using real-time data ensures it remains reliable and adaptive to evolving business conditions. The plan is to update the model with each month’s actuals on a go forward basis.
5. Developing Visualization Tools: User-friendly visualization tools can enhance stakeholder engagement and facilitate a better understanding of the model's outputs. By creating interactive and intuitive visualizations, I can enable HR, Finance, and executive teams to interact with the model's predictions and gain deeper insights into salary forecasting.
6. Ensuring Data Accuracy and Compliance: Data integrity is paramount to the success of any predictive model. Ensuring that data is up-to-date, accurate, and complies with data privacy regulations is crucial for maintaining the model's reliability and building stakeholder trust.

**Model Deployment:**

The XGBoost model emerged as the best performer in predicting future payroll expenses, showcasing its potential for supporting our financial planning and decision-making processes. Its ability to handle complex patterns and deliver accurate predictions has positioned it as an asset for our organization. However, acknowledging the dynamic nature of business environments, I recognize the importance of continuous monitoring and model updates to sustain its performance. Maintaining accuracy and relevance begins with regular monitoring of the model's outputs against actual payroll data. This ongoing evaluation allows us to assess the model's performance in real-world scenarios, identify areas of improvement, and detect potential drift in predictions. By comparing predictions with actuals on an ongoing basis, I can promptly address any deviations or discrepancies that may arise.

Additionally, I must stay vigilant about changes in internal and external factors that might influence salary expenses. Economic fluctuations, shifts in market dynamics, and evolving organizational strategies can all impact payroll patterns. Therefore, incorporating these changing factors into the model through regular updates ensures its adaptability and continued effectiveness. Collaboration among various teams, including data scientists, finance, and HR professionals, is essential for gathering valuable insights and domain expertise. By fostering open communication and collaboration, I can better align the model with business objectives and refine its features based on the changing needs of the organization.

As I gain access to new data and historical trends, incorporating the latest information into the model becomes essential for maintaining its relevance. Timely updates empower the model to capture evolving patterns and adapt to emerging salary trends, ensuring its accuracy and practicality. Data quality assurance is equally vital to safeguard the model's performance. Regularly auditing data sources and addressing data inconsistencies are integral to preserving the model's reliability. Ensuring data accuracy and compliance with privacy regulations maintains the model's credibility and instills confidence in its outputs. Ultimately, the commitment to continuous model monitoring and updates reinforces our data-driven decision-making. By keeping the XGBoost model relevant and effective, I can equip our organization with a powerful tool for salary forecasting, enabling us to navigate the future with confidence and make strategic choices that drive sustainable growth and success.

When considering the deployment strategy for the model, there are several avenues, each offering distinct advantages. One option involves the seamless integration of the model within well-established platforms like DataRobot (in evaluation stage with current employer), Azure, or AWS. Leveraging the robust infrastructure and capabilities of these platforms can facilitate a streamlined and efficient deployment process, ensuring that the model is readily accessible and effectively operationalized within our organization's existing technological ecosystem.

**Data Privacy and Protection:**

Ensuring data privacy and protection is of upmost importance as the data used for this project is private. To safeguard sensitive information, I took meticulous measures by using an identifier outside of the employee's name to attribute features to respective business units. This approach enables me to maintain the anonymity of individual employees while still capturing valuable insights for my analysis. By prioritizing data privacy and employing robust data protection protocols, I upheld the integrity and security of the data, ensuring its compliance with privacy regulations and building trust in our data-driven initiatives.

**Ethical Implications in Production:**

If this project were to be deployed in a live production environment, it is crucial to be mindful of the ethical implications that may arise concerning employee compensation, fairness, and privacy. The model's outputs should be treated with caution and subject to thorough interpretation and review to ensure they do not inadvertently introduce biases or perpetuate unfair practices in salary decisions. Transparent and accountable processes must be established to address any concerns related to fairness and equity in compensation. Preserving employee privacy and confidentiality is paramount, necessitating stringent data protection measures to safeguard sensitive information. Regular audits and oversight mechanisms should be implemented to monitor the model's performance and assess its impact on employee compensation. By embracing a strong ethical framework, I can instill confidence in the model's fairness, protect employee rights, and demonstrate our commitment to responsible and equitable use of data-driven insights in salary management.

**Mitigating Ethical Concerns:**

To effectively mitigate ethical concerns surrounding this project, a strong emphasis on transparency is essential across various aspects of model development, data usage, and decision-making processes. Transparency fosters trust and confidence among stakeholders, reassuring them that the model operates in a fair and unbiased manner. To achieve transparency, clear documentation of the model development process, including the selection of features, algorithms, and evaluation metrics, should be maintained. By making this information accessible, stakeholders can better understand how the model arrives at its predictions and assess its fairness.

Regular audits play a crucial role in ensuring ongoing compliance with ethical standards. These audits should be conducted periodically to monitor the model's performance and identify any potential biases that may arise over time. Sensitivity analysis can be employed to evaluate the model's response to changes in input data, thereby shedding light on its susceptibility to biases or inaccuracies. To address any identified biases or unfair practices, corrective measures should be promptly taken. This may involve retraining the model with updated data, refining feature selection, or adjusting algorithms to enhance fairness and equity in salary predictions. Engaging external auditors or independent review boards can provide an unbiased assessment of the model's ethical compliance. Their expertise and unbiased evaluation contribute to a more robust and trustworthy process.

Promoting diversity within the teams involved in model development and decision-making processes can aid in uncovering potential blind spots and biases. Diverse perspectives enrich the analysis and help ensure a more inclusive and ethical approach. Above all, maintaining open channels of communication and engaging in discussions with employees and stakeholders about the model's implementation and its potential impact on salary decisions is crucial. Soliciting feedback and actively addressing concerns demonstrate a commitment to transparency and ethical practices.

By integrating transparency, regular audits, sensitivity analysis, and inclusive decision-making, we can build a responsible and ethically sound framework for salary predictions. This framework fosters fairness, protects employee rights, and ensures that data-driven insights are harnessed in a manner that aligns with our organization's core values and ethical principles.

**References**

1. BIFC Corporation (2023). Payroll Data (Version 2.0) [Data set]. ADP (Automatic Data Processing).
2. Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Wiley.