Project Proposal

Title: Predicting Employee Turnover Risk at Salifort Motors

Overview:

Salifort Motors is experiencing increased employee turnover, leading to higher recruitment and training costs, reduced productivity and loss of institutional knowledge. This project will leverage data science methodologies to analyze employee data, identify key factors influencing turnover and develop a predictive model to assess employee turnover risk. The results of this analysis will provide actionable insights for Salifort Motors to proactively address employee retention, optimize resource allocation and improve overall organizational effectiveness.

Milestones	Tasks	PACE Stages
1.	Project Scoping and Data Acquisition	Plan
2.	Exploratory Data Analysis and Data Preparation	Plan, Analyze
3.	Feature Engineering and Model Selection	Analyze, Construct
4.	Predictive Model Construction	Construct
5.	Model Validation and Refinement	Analyze, Construct
6.	Turnover Risk Assessment and Analysis	Analyze
7.	Recommendations and Strategy Formulation	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of Data Science

Who is your audience for this project?

- The primary audience for this project includes Salifort Motors' executive leadership, HR department, and relevant stakeholders involved in talent management and organizational development.

What are you trying to solve or accomplish? And what do you anticipate the impact of this work will be on the larger business need?

- This project aims to:

I. Identify the primary drivers of employee turnover at Salifort Motors.
II. Develop a predictive model to quantify employee turnover risk.

III. Generate data-driven recommendations to reduce employee turnover and improve employee retention.

IV. The anticipated impact of this work is to:

- ✓ Decrease costs associated with employee turnover (e.g., recruitment, hiring, training).
- ✓ Increase employee morale and productivity.
- ✓ Enhance the effectiveness of employee retention strategies.

What questions need to be asked or answered?

- What are the key employee attributes and behaviors that correlate with turnover?
- How accurately can a predictive model identify employees at high risk of leaving?
- What specific interventions can Salifort Motors implement to mitigate turnover risk?

What resources are required to complete this project?

- Employee data from Salifort Motors' HR information system (HRIS), including demographic information, performance evaluations, compensation details, tenure, and employee survey data.
- Software and libraries for data analysis and modeling, such as Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn).

What are the deliverables that will need to be created over the course of this project?

- A comprehensive report detailing the project methodology, findings, and recommendations.
- Interactive dashboards visualizing employee turnover risk and key contributing factors.
- A predictive model for employee turnover risk.
- A presentation summarizing the project for stakeholders.

Get Started with Python

How can you best prepare to understand and organize the provided information?

- By conducting a thorough review of the HRIS data dictionary, performing initial data profiling to understand data types and distributions, and developing a data cleaning and preprocessing plan.

What follow-along and self-review codebooks will help you perform this work?

- Specify relevant Python tutorials, data manipulation guides and machine learning examples.

What are a couple of additional activities a resourceful learner would perform before starting to code?

- Researching best practices in employee turnover analysis, exploring relevant case studies, and identifying appropriate Python libraries for the project.

Go Beyond the Numbers: Translate Data into Insights

What are the data columns and variables and which ones are most relevant to your deliverable?

- Potentially relevant variables include employee demographics such as age, gender, department, job-related factors such as tenure, job role, salary, performance indicators and engagement metrics.

What units are your variables in?

- Specify units where applicable such as salary in USD, tenure in years, performance ratings on a defined scale.

What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

- Initial presumptions may include that lower job satisfaction, inadequate compensation and limited career growth opportunities are associated with higher turnover.

Is there any missing or incomplete data? Are all pieces of this dataset in the same format?

- An initial assessment of the data is required to identify missing values, inconsistencies in data formats, and potential data quality issues.

Which EDA practices will be required to begin this project?

- Descriptive statistics, data visualization such as histograms, box plots, scatter plots and correlation analysis will be employed during the exploratory data analysis phase.

- **The Power of Statistics**

What is the main purpose of this project?

- To utilize statistical methods to identify significant relationships between employee attributes and turnover and to quantify the strength of these relationships.

What is your research question for this project?

- What are the key statistical predictors of employee turnover at Salifort Motors?

What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn't use random sampling?

- Random sampling is crucial to ensure that the sample data is representative of the overall employee population, avoiding biased results.
- Sampling bias could occur if, for example the analysis only includes data from employees in a specific department or with a certain tenure level, leading to skewed conclusions.

Regression Analysis: Simplify Complex Data Relationships

Who are your stakeholders for this project?

- Key stakeholders include the HR department, management and relevant decision-makers at Salifort Motors.

What are you trying to solve or accomplish?

- To model the relationships between multiple employee variables and the likelihood of turnover using regression analysis, providing insights into the relative importance of different factors.

What are your initial observations when you explore the data?

- Record initial observations about potential trends and relationships in the data.

What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

- List relevant statistical software documentation, tutorials on regression analysis, and academic resources.

Do you have any ethical considerations in this stage?

- Ethical considerations include ensuring data privacy, avoiding the use of discriminatory variables in the model and clearly communicating the model's limitations.

The Nuts and Bolts of Machine Learning

What am I trying to solve?

- To develop a machine learning model that accurately predicts the probability of employee turnover for individual employees.

What resources do you find yourself using as you complete this stage? Is my data reliable?

- List machine learning libraries, documentation, and data quality assessment resources. Address any concerns about data reliability.

Do you have any additional ethical considerations in this stage?

- Additional ethical considerations include evaluating the fairness and potential bias of the model and establishing guidelines for the responsible use of model predictions.

What data do I need/would I like to see in a perfect world to answer this question?

- Ideally, data would include detailed information on employee reasons for leaving, employee feedback on retention initiatives and external market data on employee mobility.

What data do I have/can I get?

- Available data includes employee demographic data, performance records, compensation data, and employee survey data.

What metric should I use to evaluate success of my business objective? Why?

- Appropriate metrics include AUC-ROC to assess the model's ability to discriminate between employees who stay and those who leave and recall to minimize false negatives such as failing to identify employees who are likely to leave.

PACE: Analyze Stage

Get Started with Python

Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

- Assess data sufficiency after initial exploration and consider potential data augmentation strategies.

Go Beyond the Numbers: Translate Data into Insights

What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

- EDA will involve detailed data cleaning, feature selection, and the creation of visualizations to explore relationships between employee attributes and turnover.

Do you need to add more data using the EDA practice of joining?

- Determine whether external data sources could enhance the analysis.

What type of structuring needs to be done to this dataset such as filtering, sorting, etc.?

- Data structuring may involve filtering employees by department, sorting by tenure and aggregating data to identify trends across employee groups.

What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

- Visualizations should be clear, concise and tailored to the audience, highlighting key turnover drivers and trends.

The Power of Statistics

Why are descriptive statistics useful?

- Descriptive statistics provide a summary of the data, allowing for a better understanding of variable distributions and identification of outliers.

What is the difference between the null hypothesis and the alternative hypothesis?

- The null hypothesis assumes no significant relationship between employee factors and turnover, while the alternative hypothesis suggests a statistically significant relationship.

Regression Analysis: Simplify Complex Data Relationships

What are some purposes of EDA before constructing a multiple linear regression model?

- EDA is essential to check for linearity, multicollinearity and influential outliers, ensuring that the data meets the assumptions of regression analysis.

Do you have any ethical considerations in this stage?

- Ethical considerations include ensuring that EDA does not reveal sensitive information and that visualizations are not misleading.

The Nuts and Bolts of Machine Learning

What am I trying to solve?

- To refine the feature set and select appropriate machine learning algorithms based on insights gained from EDA.

Does it still work? Does the plan need revising? Does the data break the assumptions of the model?

- Assess the initial project plan based on EDA findings and revise if necessary. Evaluate whether the data violates modeling assumptions.

Why did you select the X variables you did? What are some purposes of EDA before constructing a model? What has the EDA told you?

- Explain the rationale for selecting predictor variables based on EDA and domain knowledge.

What resources do you find yourself using as you complete this stage?

- Document resources used for EDA, feature engineering and model selection.

Do you have any ethical considerations in this stage?

- Evaluate the fairness of selected features and the potential for discriminatory bias in the model.

PACE: Construct Stage

Get Started with Python

Do any data variables averages look unusual?

- Identify and investigate any outliers or anomalies in the data.

How many vendors, organizations or groupings are included in this total data?

- Determine the scope of the data, such as the number of departments or business units.

Go Beyond the Numbers: Translate Data into Insights

What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

- Develop a classification model to predict turnover risk, create visualizations to communicate key drivers of turnover and generate reports summarizing findings.

What processes need to be performed in order to build the necessary data visualizations?

- Data aggregation, chart type selection and formatting to ensure effective communication of insights.

Which variables are most applicable for the visualizations in this data project?

- Variables that highlight significant differences between employees who stay and those who leave and those that reveal key trends in turnover.

Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

- Address missing data using appropriate imputation or removal techniques as determined during the data preparation phase.

The Power of Statistics

How did you formulate your null hypothesis and alternative hypothesis?

- **Null Hypothesis**: There is no statistically significant relationship between the selected employee factors and employee turnover.
- **Alternative Hypothesis:** There is a statistically significant relationship between the selected employee factors and employee turnover.

What conclusion can be drawn from the hypothesis test?

- State the outcome of the hypothesis test and its implications.

Regression Analysis: Simplify Complex Data Relationships

Do you notice anything odd? Can you improve it? Is there anything you would change about the model?

- Identify any issues with the regression model and suggest potential improvements.

The Nuts and Bolts of Machine Learning

Is there a problem? Can it be fixed? If so, how?

- Address any problems with the machine learning model, such as overfitting or underfitting.

Which independent variables did you choose for the model, and why?

- Justify the selection of predictor variables based on EDA and domain knowledge.

How well does your model fit the data? (What is my model's validation score?)

- Report relevant model performance metrics.

Can you improve it? Is there anything you would change about the model?

- Suggest potential model refinements.

Do you have any ethical considerations in this stage?

- Ensure model fairness and address potential for discriminatory outcomes.

PACE: Execute Stage

Get Started with Python

Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?

- Suggest specific areas for further investigation based on initial data insights.

What data initially presents as containing anomalies?

- Identify any data points that appear unusual or potentially erroneous.
- What additional types of data could strengthen this dataset?

Additional data, such as employee exit interview feedback and external market data, could enhance the analysis.

Go Beyond the Numbers: Translate Data into Insights

What key insights emerged from your EDA and visualizations(s)?

- Summarize the main findings from the exploratory data analysis.

What business recommendations do you propose based on the visualization(s) built?

- Provide actionable recommendations to Salifort Motors based on the data analysis.

Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

- Suggest further research questions that could be explored to deepen the understanding of employee turnover.

How might you share these visualizations with different audiences?

- Describe how visualizations would be tailored for different stakeholders such as HR, management, and employees.

The Power of Statistics

What key business insight(s) emerged from your A/B test?

- Summarize the findings from any hypothesis tests conducted.

What business recommendations do you propose based on your results?

- Provide recommendations based on the statistical analysis.

Regression Analysis: Simplify Complex Data Relationships

To interpret model results, why is it important to interpret the beta coefficients?

- Interpreting beta coefficients is crucial for understanding the direction and magnitude of the relationship between predictor variables and employee turnover.