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

Predicting Employee Turnover to Improve Retention

Project Proposal

Overview:

This project aims to develop a machine learning model to predict employee turnover within a company. By identifying factors that contribute to employees leaving, the company can proactively implement strategies to improve retention. The project involves exploratory data analysis (EDA), statistical analysis, and the development and evaluation of predictive models

Milestones	Tasks	PACE Stages
Data Understanding & Preparation	Load and inspect data, handle missing values and duplicates, rename columns, identify outliers.	Plan, Analyze
Exploratory Data Analysis (EDA)	Visualize key variables, analyze distributions, identify correlations, explore relationships between variables and turnover.	Analyze, Construct
Statistical Analysis	Compute descriptive statistics, conduct correlation analysis (Point-biserial, Cramer's V), formulate and test hypotheses	Analyze, Construct
Model Building & Evaluation	Preprocess data (encoding, scaling), split data, build and train Logistic Regression and Random Forest models, evaluate model performance using various metrics (accuracy, precision, recall, F1, ROC AUC), analyze feature importance.	Construct, Execute
Insights & Recommendations	Interpret model results, identify key drivers of turnover, formulate actionable business recommendations, consider ethical implications.	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- **Who is your audience for this project?** HR managers, executive leadership, and department heads.
- **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?** I am building a model to identify employees likely to leave. This work aims to reduce turnover, saving recruitment costs and retaining knowledge.
- **What questions need to be asked or answered?**
    - What drives employee turnover?
    - Can we predict which employees will leave?
    - What insights can improve employee retention?
    - Are certain departments or salary levels experiencing higher turnover?
- **What resources are required to complete this project?**
    - HR capstone dataset.
    - Python with pandas, numpy, seaborn, matplotlib, scipy, sklearn, imblearn, xgboost, shap.
    - Jupyter Notebook.
    - Tableau (if needed).
- **What are the deliverables that will need to be created over the course of this project?**
    - A Jupyter Notebook with code, analysis, and visualizations.
    - An executive summary of findings and recommendations.
    - A presentation deck (optional).

Get Started with Python

- **How can you best prepare to understand and organize the provided information?** Load data into pandas, use df.info() and df.describe().T for an overview, and rename columns.
- **What follow-along and self-review codebooks will help you perform this work?** “Foundations of Data”, “Regression Analysis” and “Nuts and Bolt of Machine Learning”
- **What are a couple additional activities a resourceful learner would perform before starting to code?** Research common HR analytics problems and review similar employee turnover projects.

Go Beyond the Numbers: Translate Data into Insights

- **What are the data columns and variables and which ones are most relevant to your deliverable?** satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, work_accident, promotion_last_5years, department, salary, and left (target). All are relevant.

- **What units are your variables in?** satisfaction_level float64, last_evaluation float64, number_project int64, average_monthly_hours int64, time_spend_company int64, work_accident int64, left int64, promotion_last_5years int64, department object, salary object.
- **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?** Lower satisfaction, high hours, and longer tenure might correlate with turnover. Lack of promotion and accidents could also play a role.
- **Is there any missing or incomplete data?** I will check using df.isna().sum().
- **Are all pieces of this dataset in the same format?** I will verify data types and plan for encoding categorical variables.
- **Which EDA practices will be required to begin this project?** Descriptive statistics, checking for missing values/duplicates, outlier detection, correlation analysis, and various visualizations.

The Power of Statistics

- **What is the main purpose of this project?** To build a predictive model for employee turnover and identify key factors for retention.
- **What is your research question for this project?** Can we predict employee turnover from their data, and what are the most significant predictors?
- **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 ensures the sample represents the population. Not using it could lead to bias, ex: sampling only one department.

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?** Yes, the data appears sufficient for turnover prediction.

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?** Understand data structure, clean data (duplicates/outliers), perform univariate and bivariate analysis, and identify key insights.
- **Do you need to add more data using the EDA practice of joining?** No, the current dataset is sufficient for initial modelling.
- **What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?** Outliers were filtered, salary (ordinal) and department (one-hot) will be encoded.
- **What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?** High-level charts for executives, detailed box plots, heatmaps, and SHAP plots for analysts.

The Power of Statistics

- **Why are descriptive statistics useful?** Descriptive statistics provide data summaries (mean, median, std dev) that help identify outliers, understand distributions, and inform preprocessing decisions
- **What is the difference between the null hypothesis and the alternative hypothesis?** H_0 states no effect/relationship, H_1 (alternative) states there is an effect/relationship.

Regression Analysis: Simplify Complex Data Relationships

- **What are some purposes of EDA before constructing a multiple linear regression model?** Identify relationships, check for multicollinearity, detect outliers, assess assumptions, and prepare categorical variables.
- **Do you have any ethical considerations in this stage?** Yes: address data bias, ensure privacy, avoid implying causation, and accurately represent findings.

The Nuts and Bolts of Machine Learning

- **What am I trying to solve?** To classify employees as likely to stay or leave.
- **Does it still work? Does the plan need revising?** The plan is good, data fits classification.
- **Does the data break the assumptions of the model? Is that ok, or unacceptable?** Logistic Regression has assumptions (linearity, independence, no multicollinearity), Random Forest is more robust. Violations for Logistic Regression may reduce reliability.
- **Why did you select the X variables you did?** All independent variables were chosen as they are potentially relevant to turnover.
- **What are some purposes of EDA before constructing a model?** (Same as above)
- **What has the EDA told you?** Duplicates and outliers exist, target variable (left) is imbalanced (3.08:1 stayed:left), satisfaction_level strongly correlates negatively with turnover.
- **What resources do you find yourself using as you complete this stage?** Pandas, seaborn, matplotlib, and scikit-learn.
- **Do you have any ethical considerations in this stage?** (Same as above)

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- **Do any data variables averages look unusual?** No, but time_spend_company had outliers which were handled.
- **How many vendors, organizations or groupings are included in this total data?** Multiple departments and salary levels are included.

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?** Box plots, correlation

heatmaps, ridge plots, stacked bar charts, turnover rate heatmaps, Logistic Regression and Random Forest models, evaluation metrics, confusion matrices, and feature importance plots.

- **What processes need to be performed in order to build the necessary data visualizations?** Data cleaning, aggregation, and plotting using matplotlib.pyplot and seaborn.
- **Which variables are most applicable for the visualizations in this data project?** Key numerical variables for box plots/heatmaps, categorical variables for bar charts.
- **Going back to the Plan stage, how do you plan to deal with the missing data (if any)?** No missing data was found.

The Power of Statistics

- **How did you formulate your null hypothesis and alternative hypothesis?** For correlation, H_0: no correlation ($\rho=0$), H_1: correlation ($\rho\neq0$). For independence, H_0: independent, H_1: dependent.
- **What conclusion can be drawn from the hypothesis test?** Significant negative correlation between satisfaction_level and turnover, weak but significant positive correlation between monthly_hours and turnover, weak association between department and turnover.

Regression Analysis: Simplify Complex Data Relationships

- **Do you notice anything odd?** I will monitor for unexpected coefficient signs/magnitudes.
- **Can you improve it? Is there anything you would change about the model?** Yes, further improvements will be considered in the Execute stage.

The Nuts and Bolts of Machine Learning

- **Is there a problem? Can it be fixed? If so, how?** Class imbalance is a problem, addressed by class_weight='balanced' or SMOTE.
- **Which independent variables did you choose for the model, and why?** All independent variables were chosen as they are potentially relevant to predicting turnover.
- **How well does your model fit the data? (What is my model's validation score?)** Random Forest (Accuracy: 0.9909, ROC AUC: 0.9911) shows excellent fit. Logistic Regression (Accuracy: 0.8156, ROC AUC: 0.8793) also performed well.
- **Can you improve it? Is there anything you would change about the model?** No
- **Do you have any ethical considerations in this stage?** No

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?** A preliminary check for data completeness, consistency, and clear definitions of columns.

- **What data initially presents as containing anomalies?** time_spend_company had outliers.
- **What additional types of data could strengthen this dataset?** Employee feedback/surveys, manager effectiveness ratings, training opportunities, competitor salary benchmarks, and anonymized demographics.

Go Beyond the Numbers: Translate Data into Insights

- **What key insights emerged from your EDA and visualizations(s)?** 24.5% turnover, satisfaction_level strongly impacts turnover, employees who left had lower satisfaction, higher evaluation/hours/tenure, department/salary impact turnover.
- **What business recommendations do you propose based on the visualization(s) built?** Address satisfaction, monitor workload, review evaluations/promotions, and target retention efforts to high-turnover areas.
- **Given what you know about the data and the visualizations you were using, what other questions could you research for the team?** Specific causes of low satisfaction, thresholds for high hours, manager quality impact, project-specific satisfaction.
- **How might you share these visualizations with different audiences?** Executive summary with key charts, detailed visualizations for HR, department-specific views for department heads, interactive dashboards for broader access.

The Power of Statistics

- **What key business insight(s) emerged from your A/B test?** Based on correlation analysis, satisfaction_level is a critical factor, average_monthly_hours and department also show significant associations.
- **What business recommendations do you propose based on your results?** Prioritize satisfaction initiatives, manage workloads, and analyze department-specific turnover issues.

Regression Analysis: Simplify Complex Data Relationships

- **To interpret model results, why is it important to interpret the beta coefficients?** For Logistic Regression, beta coefficients (or odds ratios) show the impact of each variable on turnover likelihood. For Random Forest, feature importance/SHAP values provide similar insights.
- **What potential recommendations would you make to your manager/company? Do you think your model could be improved? Why or why not? How?**
    - **Recommendations:** Implement retention programs, work-life balance initiatives, performance management review, department-specific strategies.
    - **Improvement:** Yes, through hyperparameter tuning, other ensemble methods (XGBoost), feature engineering, or collecting more data.
- **What business recommendations do you propose based on the models built?** Focus on satisfaction_level, time_spend_company, number_project, average_monthly_hours, and last_evaluation for retention strategies.
- **What key insights emerged from your model(s)?** Random Forest is highly accurate, satisfaction_level is the most important feature, followed by time_spend_company, number_project, average_monthly_hours, and last_evaluation.

- **Do you have any ethical considerations at this stage?** Yes: responsible model use, transparency, and bias mitigation.

The Nuts and Bolts of Machine Learning

- **What key insights emerged from your model(s)?** (Same as above)
- **What are the criteria for model selection?** Predictive performance (ROC AUC, F1-score), interpretability, robustness, and business relevance.
- **Does my model make sense? Are my final results acceptable?** Yes, the Random Forest model's high performance and intuitive feature importance make results acceptable.
- **Were there any features that were not important at all? What if you take them out?** Less important features exist. Removing them can simplify the model and reduce training time without significant performance loss.
- **Given what you know about the data and the models you were using, what other questions could you address for the team?** Financial impact of turnover reduction, early warning systems, and segment-specific turnover drivers.
- **What resources do you find yourself using as you complete this stage?** Scikit-learn for evaluation/interpretation, and SHAP for explainability.
- **Is my model ethical?** Yes, it predicts turnover using work-related factors for retention strategies, maintaining privacy.
- **When my model makes a mistake, what is happening? How does that translate to my use case?**
    - **False Positives:** Model predicts turnover, but employee stays. Retention resources might be misallocated.
    - **False Negatives:** Model predicts stay, but employee leaves. This is a critical error, as a retention opportunity is lost. High recall minimizes this.