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

Salifort Motors Employee Turnover Analysis Project Proposal

Overview

Salifort Motors is experiencing a high amount of employee turnover. The goal is to find a hidden reason or reasons in the employee data as to what is causing this issue. The use of exploratory data analysis, statistical analysis, and machine learning models will be key to understanding the data features that are correlated with the turnover.

Milestones	Tasks	PACE stages
1	Write proposal and gather information and data	Plan
2	Exploring and cleaning data	Plan and Analyze
2a	Compute descriptive statistics and build visualizations	Analyze and Construct
3	Conduct hypothesis testing	Analyze and Construct
4	Build machine learning models	Analyze and Construct
4a	Evaluate models	Execute
5	Communicate final insights to stakeholders and present final model	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?
    - The executives of Salifort are the audience for the project. They will be receiving any final product to be used for determining whether an employee will leave. Additionally, they are looking to understand why the turnover rate is so high, whether it can be linked to a department or some other data feature.
- 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?
    - We are looking to determine why the company's turnover rate is so high and whether we can build a model that will be able to predict an employee leaving the company.
- What questions need to be asked or answered?
    - Is there a department that is a clear indicator of leaving?
    - Are there incentives that keep employees longer?
    - Will a machine learning model be able to find a pattern in the data to correctly predict turnover?
- What resources are required to complete this project?
    - The resources required for this project are computational and the employee data.
- What are the deliverables that will need to be created over the course of this project?
    - The deliverables for this project are going to be visualizations, hypothesis testing results, and a machine learning model if actionable data patterns are found.

Get Started with Python

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

  - We can become best acquainted and understand the data provided through reading the data dictionary to understand the contents of each column and to run pd.DataFrame.info() and .describe(). These will give us information on the number of null values, descriptive statistics, and general distribution of the data set.
- What follow-along and self-review codebooks will help you perform this work?
  - The use of the documentation of the Pandas library helped with any questions with respect to data preparation and organization
- What are a couple additional activities a resourceful learner would perform before starting to code?
  - At this point we can take the time to review the features and do feature transformations, incorporating features into metrics that could tell a better story of the data for the visual and statistical review as well as for the model's training and interpretation.

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, average_monthly_hours, time_spend_company, work_accident, promotion_last_5years, salary, contribution_rate, department
- What units are your variables in?
  - These variables include the units of: hours, years, count, categorical binary, dollars, count per year, categorical department. Through processing they will all be changed to a numeric unit, such as int64, including the categorical variables, so that the models can better handle them.
- 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?
  - There is a large amount of turnover from three departments; however, those are the largest departments and appear to be roughly the same scale of turnover when comparing department size to turnover by department. The data is slightly skewed with only approximately 16% of 'left' being turnover. We might need to think about up/down sampling when it comes to modeling.
  - My initial presumption is there definitely appears to be relationships between 'left' and a number of the variables, there also doesn't appear to be much collinearity.
- Is there any missing or incomplete data?
  - There is nothing blatantly incomplete or missing. However, there were a large number of duplicate rows. It is not clear as to whether this was a coincidence and different employees have the identical stats.

- Are all pieces of this dataset in the same format?
    - Some pieces of the dataset are in the unit of object the rest are numeric. There is one column (average hours per month) that is in the hundreds while the rest are in the tens are less.
- Which EDA practices will be required to begin this project?
    - Prior to beginning EDA, we must check for duplicates and null values. During EDA we can check for outliers and deal with them as it seems fit.

The Power of Statistics

- What is the main purpose of this project?
    - The main purpose of this project is to predict what employees will turnover in the company and hopefully understand which features of the company are driving this high turnover rate.
- What is your research question for this project?
    - Research question for this project could be: is the turnover for Salifort different from the turnover of other companies. Another question that is much more feasible to answer is whether there is a difference in the turnover of high versus low satisfaction reporting employees?
- 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?
    - The use of random sampling is important since without it we would get a non-comprehensive profile of the company's employees. This could skew the projections from analysis. We are already seeing an imbalance of the number of turnover employees and this should be reflected in the sampled data. An additional note is that the HR collected data via survey, which relies heavily on the taker telling the truth and that the taker returns on the data.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
    - The executives of Salifort are the audience for the project. They will be receiving any final product to be used for determining whether an employee will leave. Additionally, they are looking to understand why the turnover rate is so high, whether it can be linked to a department or some other data feature.

- What are you trying to solve or accomplish?
    - We are looking to determine why the company's turnover rate is so high and whether we can build a model that will be able to predict an employee leaving the company.
- What are your initial observations when you explore the data?
    - Initial observations of the data. There is one column of a different scale than the others, average_monthly_hours is in the hundreds, while the rest are between 0 and 1 and on the tens scale. There are two categorical variables: Department and salary that will need to be handled. Our target variable, "left" , is already in a binary categorical encoding.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
    - The resources required for this project are computational and the employee data. The libraries that we are using are Pandas: https://pandas.pydata.org/docs/index.html Numpy: https://numpy.org/ Seaborn: https://seaborn.pydata.org/ Matplotlib: https://matplotlib.org/ Scikit Learn: https://scikit-learn.org/stable/, SciPy: https://docs.scipy.org/doc/scipy/index.html
- Do you have any ethical considerations in this stage?
    - At this stage the ethical considerations would be data privacy and permission for use of personal information, that the data and insights be used fairly and responsibly, and is there bias being observed and or introduced into the modeling. I would assume that the employee data is owned by the company and the employees have signed their permission to this data being recorded. The data is owned by the company and questions being answered are for the company, so we must assume that the executives using the insights and model made here will use it fairly and ethically. This is the perfect time to hide age, gender, and race if they are part of the data, so that any trends related to them while reviewing the data and building models will not be formative to the model's learning or any observations made.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
    - I am trying to solve whether we can predict if an employee will be part of the turnover.
- What resources do you find yourself using as you complete this stage?
    - At this point I used confusion matrix and metrics tools from scikit learn and its supporting documentation, as well as seaborn, matplotlib, pandas.
- Is my data reliable?

    - The data is a primary data source. It is empirical data that has been collected from the HR survey. The data is flawed and can introduce bias into the data since, people more inclined to reply may do so and they might lie on certain stats, especially one such as satisfaction level.
- Do you have any additional ethical considerations in this stage?
    - At this point the concern is about potential bias being trained into the model and then causing skewed incorrect information being shared to the executives of the company and all stakeholders.
- What data do I need/would I like to see in a perfect world to answer this question?
    - I would like to see observed data , not reported, that can be analyzed without worrying about the consideration that the employees may be not telling the full truth with respect to their satisfaction, hours, etc.
- What data do I have/can I get?
    - This data is reported by employees, and it can likely be cross checked with company records about project involvement, employment duration, works worked, etc. We can also build new features such as “contribution_rate” from the current features.
- What metric should I use to evaluate success of my business objective? Why?
    - The metric that should be used in this case is F1, since it helps strike a balance between Precision and Recall. Additionally, this is helpful when the dataset is imbalanced, which once the data set is cleaned, approximately 16% of the employees are involved in turnover and that is below the 20% rule of thumb.

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?
    - There is a large data set, while being slightly imbalanced any patterns related to the target variable should be modelable.

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?
    - Check for outliers, look for linear relationships, check for duplicates, check for null values. Handle all issues as needed, This is a classification problem, and I am not seeing any linear relationships or collinearity, I am guessing I will be using a Decision Tree, Random Forest, or XGBoost and do not need to worry about outliers.
- 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.?
    - No additional data needed to be added. However, I did concatenate result tables to better compare all of the results of the models as they were being built.
- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
    - Our intended audience is the executives of the company. I need to show distinctly through visualizations what features are showing patterns related to the ‘left’ target variable. This will include bar plots, histograms, and scatter plots, while clearly indicating the data points of turnover employees.

The Power of Statistics

- Why are descriptive statistics useful?
    - Descriptive statistics are useful since they help initially visualize the shape of the data and whether there are large amounts of outliers and issues with scale of the data points.

- What is the difference between the null hypothesis and the alternative hypothesis?
    - The null hypothesis is the case where there is no change or difference in the data of concern versus random chance. While the alternative hypothesis is that there is a statistically significant difference in the data, such that we must reject that the null hypothesis or no change/no difference is true.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
    - The purpose for EDA before constructing a multiple linear regression model is to check for a linear relationship between the target variable and the associated features as well as to check for no collinearity between features.
- Do you have any ethical considerations in this stage?
    - At this stage the ethical considerations would be data privacy and permission for us of personal information, that the data and insights be used fairly and responsibly, and is there bias being observed and or introduced into the modeling

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
    - We are trying to understand why the turnover is high and predict what employees will be part of it. There are no linear relationships so other models need to be approached such as Decision Tree.
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
    - The data breaks the assumptions for linear regression, since there isn't any linear relationship between features and target and there are outliers, but these both do not not matter with respect to Decision Tree, Random Forest, etc.
- Why did you select the X variables you did?
    - I selected the X variables since they all could be a factor influencing turnover and could be the one of the features heavily causing turnover, determining which is part of our project goal.
- What are some purposes of EDA before constructing a model?
    - The purpose of EDA before modeling allows us to understand the type of models we should try, and whether we need to further modify the dataset.
- What has the EDA told you?

- EDA has shown that there are no linear relationships. However, it has shown that there are many factors with relationships to the 'left' target variable. For instance, the majority cluster of employees that were turnover had the contribution rate between 1 and 2 projects/year.

- What resources do you find yourself using as you complete this stage?
  - During this phase I used matplotlib, pandas, and python. referencing the documentation provided by their authors.
- Do you have any ethical considerations in this stage?
  - At this point if there were any demographic information it should be hidden prior to moving on to modeling. Data privacy should always be a concern. Additionally, I must be clear with the stakeholders how I have manipulated the data so as to not mislead anyone from the truth.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
    - The average number of hours per week is very high at 201 hours, indicating that many employees are working well over a 40hr week.
- How many vendors, organizations or groupings are included in this total data?
    - The data only includes data of employees surveyed inside Salifort Motors.

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?
    - The data visualizations include many different comparisons of the relationship of 'left' and the features. After getting any idea what is impacting turnover on a feature level, then we will produce a machine learning model to predict which employees will turnover.
- What processes need to be performed in order to build the necessary data visualizations?
    - Process for data visualizations: access and clean data, group and sort data into sub variables, build figure in seaborn/matplotlib.
- Which variables are most applicable for the visualizations in this data project?
    - Comparing many of the features to the 'left' variable in barplot, scatter plots, and histograms reveal the relationships. I found the satisfaction score histogram to be a very telling visualization. We see there that there is a group of 'turnover' employees with high satisfaction and low satisfaction. This indicates that there are at least two problems affecting the company, since a group of employees are feeling good about their work and then still leaving, same with a group that is not feeling good about their work.
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

- There is a slight imbalance in the distribution of the data set when I drop the duplicated rows. I have decided to test both the data including and dropping the duplicates and see which models handle validation better after that training.

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
  - The question that I would love to answer most is whether the turnover for this company is different from any other company. That is data that I; however, cannot acquire.
    - Null Hypothesis: No difference in mean 'left' between employees with 'high' or 'low' satisfaction
    - Alternative Hypothesis: There is a difference in mean 'left' between employees with 'high' or 'low' satisfaction
- What conclusion can be drawn from the hypothesis test?
  - The conclusion is that we reject the Null hypothesis with a p-value of 3.136e-89 in a two-tail t-test, clearly there is a difference in mean ‘left’ between employees with high and low satisfaction.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?
  - There is no linear relationships and therefore we should not use a logistic regression machine learning model
- Can you improve it? Is there anything you would change about the model?
  - With some feature transformations then we might find a linear relationship to apply an effect regression analysis to the dataset.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?
  - The prediction of whether an employee will be turnover is be best addressed by the building decision tree, random forest, or xgboost model.
- Which independent variables did you choose for the model, and why?
  - Satisfaction_level, last_evaluation, average_monthly_hours, time_spend_company, work_accident, promotion_last_5years, salary, contribution_rate, department

    - These variables were chosen since they all could be a factor influencing turnover and could be the one of the features heavily causing turnover, determining which is part of our project goal.
- How well does your model fit the data? (What is my model's validation score?)
    - The Random Forest model had the best weighted average validation scores of 0.99 for recall, precision, f1, and accuracy.
- Can you improve it? Is there anything you would change about the model?
    - There is always potential for manipulating the hyperparameter tuning. The model trained on the data set that still included the duplicated data had a better score than other models based on best_estimator training scores; however, during validation it did not run as well, leading me to believe that it most likely started to get over fit on the training set.
- Do you have any ethical considerations in this stage?
    - At this point the continued concern about potential bias being trained into the model and then causing skewed incorrect information being shared to the executives of the company and all stakeholders.

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?
    - From what we have seen from the visuals produced, my manager and executives should immediately look into implementing standard working hours and assessing how employees are scored by their manager's review.
- What data initially presents as containing anomalies?
    - There are a high number of duplicate rows. There are outliers in the data, time_spent_company for example, however, with the Random Forest model being chosen, outliers are handled by the model itself.
- What additional types of data could strengthen this dataset?
    - Data that is empirical and does not contain the chance of being biased or untrue would really strengthen the data set. For instance, recording time for certain task completion. This would be different by department and not applicable to all employees most likely.
    - With time investment, developing and implementing a task point system company wide. This would allow for efficiency and quality monitoring and give a more unbiased data to strengthen the overall dataset.
    - Data on raises and the timing of them.

Go Beyond the Numbers: Translate Data into Insights

- What key insights emerged from your EDA and visualizations(s)?
    - Turnover is relatively the same rate over all departments
    - Employees with a contribution rate between 1 and 2 projects/year have a high turnover
    - Management has the lowest turnover rate across all pay brackets
    - Turnover increases exponentially with time at the company, peaking after 5 years and then falls off rapidly.
    - A large portion of employees that turnover have a satisfaction score below 0.5. However, there is still a lot of turnover that happens in employees with a satisfaction over 0.5. This could indicate there are at least two major internal issues pushing turnover.

  - Similar phenomenon with respect to manager review, there is a group with bad reviews and a group with good reviews. There are at least two issues driving turnover.
  - Employees with very low satisfaction scores work the most hours.
  - Work accidents seem to have a small effect on turnover
  - All employees who are turnover has not had a promotion in the last five years. This is likely a major driving force. There is little or no monetary acknowledgement to employees committed to the company.

- What business recommendations do you propose based on the visualization(s) built?
  - I would recommend looking to better enforce regular working hours, while also developing a new system for advancement in the company.
  - There is likely no reason to stay with the company around year 3 or 4 advancement-wise, thus we see a spike in turn over at year 5. Promotion/raises/incentives appear to be necessary.
- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
  - How many employees are working longer hours?
  - Does a 'medium' salary affect the rate of turnover versus 'low'?
- How might you share these visualizations with different audiences?
  - On an executive level the bar chart and histogram visualizations should be shown, since they are easily interpretable. These visualizations can also be shown to a more technical audience due to the important content shown in them. The key to bridge the gap between the general audience and technical is proper and clear definition and display of units, variables, and content.

The Power of Statistics

- What key business insight(s) emerged from your A/B test?
  - We found that there was a statistical difference between the average of 'left' for high satisfaction versus low.
- What business recommendations do you propose based on your results?
  - Management needs to find a way to improve the satisfaction of employees. It is a statistically significant driver of the turnover in the company.

Regression Analysis: Simplify Complex Data Relationships

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

    - Beta coefficients allow you to estimate the magnitude and direction (positive or negative) of the effect of each independent variable on the dependent variable. The coefficient estimates can be converted to explainable insights.

- What potential recommendations would you make to your manager/company?
    - Start increasing salary and introduce more incentives.
- Do you think your model could be improved? Why or why not? How?
    - Since there is no real linear relationship, I moved on to using machine learning models to better handle this situation.
- What business recommendations do you propose based on the models built?
    - I would recommend looking to better enforce regular working hours, while also developing a new system for advancement in the company.
    - There is likely no reason to stay with the company around year 3 or 4 advancement-wise, thus we see a spike in turn over at year 5. Promotion/raises/incentives appear to be necessary.
- What key insights emerged from your model(s)?
    - We find that the model can predict which employees are going to be turnover at a consistently high rate of confidence.
    - We observed the Random Forest model rise to the top and barely out perform the Decision Tree model. Accuracy and weight average recall, precision, and f1 are all 98%. With the testing recall of 90%, precision of 98%, and f1 of 93%.
    - 'satisfaction_level', 'time_spend_company', 'average_monthly_hours', 'contribution_rate', and 'last_evaluation' are the 5 most important features.
- Do you have any ethical considerations at this stage?
    - Continued concern about potential bias being trained into the model and then causing skewed, incorrect information being shared to the executives of the company and all stakeholders.

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?
    - I would recommend that we implement the Random Forest model and so as to better predict the employee movements through turnover. Additionally, and potentially more importantly, we find these high importance features that might be indicating acute issues within the company. Employees are dissatisfied, spending more time per month, and leaving after years of no acknowledgement.
    - I would recommend managers to address working hours and workload based issues, clearly there is upset in the work force and this is likely also affecting the evaluation scores. Therefore leading to layoffs and employees leaving.

    - Further steps would be trying to improve the model with tuning and feature transformations. I would recommend taking action with improving workplace conditions that could start improving turnover as soon as possible.
    - Looking to implement a task points work system to better monitor employee efficiency and quality of work and bring a much needed potentially unbiased data to the dataset.

- What are the criteria for model selection?
    - Since the data set is slightly imbalanced and looking for the small target variable, contains outliers, and does not have linear relationships between the target variable and the features then an ensemble learning model, such as random forest can handle it better.
    - The model is trained to select for best 'F1' which balances the precision and recall and is better for handling imbalanced datasets.
- Does my model make sense? Are my final results acceptable?
    - The model predicts the dataset with a 98% accuracy. From the test dataset, the results are: recall of 90%, precision of 98%, and F1 of 93%. I feel that these are acceptable results and will help for future predictions.
- Were there any features that were not important at all? What if you take them out?
    - Department information and promotion information seemed to have little to no impact on the model output.
    - We see the F1 score increase to 94%, all other scores remained the same.
- Given what you know about the data and the models you were using, what other questions could you address for the team?
    - What are the drivers for the high turnover in the 5th year at the company? Is the population number after 5 and 6 years extremely small?
        - We need information on the lay off habits of the company as well
    - How does salary affect turnover? Is the turnover rate different between the three tiers?
- What resources do you find yourself using as you complete this stage?
    - At this point I used confusion matrix and metrics tools from scikit learn and its supporting documentation, as well as seaborn, matplotlib, pandas.
- Is my model ethical?
    - Since the data is self reported, any biases or inaccuracies are due to any falsehoods in the data. This could also point to a company wide issue of employees being afraid to submit truthful statements.

    - So long as the model is used for the betterment of the company while keeping the privacy of individuals and their permissions safe, then the model is working within the ethics laid down by a private company doing an interview review.

- When my model makes a mistake, what is happening? How does that translate to my use case?
    - We are seeing most commonly False Negative predictions
    - False negatives are employees that will “turnover” but have been predicted to not turnover.
    - These incorrect predictions are the ones we want to avoid. I would recommend further improving the model through:
        - Different feature selection and transformation as well as hyperparameter tuning.
        - Additionally, building better unbiased metrics into the company, i.e. project points, to understand how each employee is doing would likely bring about better predictions once a new model is trained on this data.