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

Predicting Employee Longevity through Supervised Machine Learning Models

Overview

We built a hyper-tuned random forest classifier to predict whether or not an employee would leave the company based on performance, past projects, department, satisfaction level, salary level and time spent with the company.

Milestones	Tasks	PACE stages
1st	Clean data, remove outliers from continuous columns, remove rows with empty entries, and duplicated rows.	Plan
2nd	Look for correlations between data and the target variable such that we can select suitable features for future models. Assumptions such as no multicollinearity and observational independence need to be met before we proceed.	Analyze
3rd	Explore data, look at class balances for each dpt. as this could skew results if dpt. size is not considered. Visualize distributions of continuous variables for each class i.e. dpt. and salary level.	Analyze

4th	Explore if each dpt. has higher proportions of employees leaving than other departments. Compare the proportions of income levels in each department to and see if there is a relation with employee leaving.	Analyze
5th	Select features that have high correlation with target variables and are close to independent from one another. Split the data into training/validation/testing sets for cross validation.	Constructing
6th	Build the XGBooster machine, and a Random Forest Classifier and tune their hyper-parameters through GridSearch. Evaluate each model through the 4 metrics (Accuracy, Precision, Recall, F1) and select the champion model.	Construct
7th	Run the champion model on the test set, evaluate performance, look at feature importance and draw connections to previous information from EDA. Provide business suggestions based on results from champion model relevant feature behavior and correlations amongst features.	Evaluation

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?
  - The audience for this project are the management team who are looking for data driven answers towards their investments.
- 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 trying to develop a model that can predict the longevity of employees based on their activity, their department, their salary etc. such that management knows what kind of people they can safely invest on and expect to keep around for the foreseeable future.
- What questions need to be asked or answered?
  - What drives employees to leave the company?
  - What can the company do to retain more employees?
  - What departments have the highest proportion of employees leaving the company?
- What resources are required to complete this project?
  - Clearly we need past and present employee data and company data.
- What are the deliverables that will need to be created over the course of this project?
  - We will need to create multiple figures that display the results of our findings and tables with the evaluations of the model and its performance.

Get Started with Python

- How can you best prepare to understand and organize the provided information?
  - The best way to prepare and organize the information provided to us is by having a clear goal in mind such that every decision made in the process of organizing data is driven to achieve the end result.
- What follow-along and self-review codebooks will help you perform this work?

    - To perform EDA I used notebooks from the following courses: “Foundations of Data”, “Regression Analysis” and “Nuts and Bolt of Machine Learning”. This notebooks refreshed me on the usage of several functions from multiple libraries in python.
- What are a couple additional activities a resourceful learner would perform before starting to code?
    - Create a checklist of goals/milestones you need to accomplish for this project in order to fully satisfy business needs.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
    - The data columns for this project contain employee statistics such as
        - Satisfaction
        - Last Evaluation
        - Number of projects
        - Average monthly hours
        - Time spent with the company
        - Work Accidents
        - Promotions in last 5yr
        - Department
        - Salary level
    - We chose the significance of the variables by looking at their correlation factor to the target variable which in this case is whether an employee leaves the company.
    - We also select variables that are not too correlated with each other such that the independence assumption is met
- What units are your variables in?
    - There is three different types of variables in this dataset
        - Continuous variables (Satisfaction, evaluation, number projects, monthly hours, and time spent with company)
        - Categorical variables (Work accident, promotion in last 5yr, department and salary)
- 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 is that the management department will have the higher proportion of high income employers and the lowest turnover rate.
  - Part of my believes that salary, and promotions, may be important factors towards an employee leaving the company
  - There has to be a high correlation between time spent in company and promotions in the last 5yrs.
- Is there any missing or incomplete data?
  - There is no missing data.
- Are all pieces of this dataset in the same format?
  - The data types vary according to what has been previously mentioned for each column.
- Which EDA practices will be required to begin this project?
  - We need to look at the source of the data and identify and possible bias this source could introduce to our analysis and how we would solve it.

The Power of Statistics

- What is the main purpose of this project?
  - The main purpose of this project is to develop a classification model that can predict employee turnover rates given some statistics of the employee's performance and activities.
- What is your research question for this project?
  - What features within the data-set provided are good indicators of employee retention for this company?
- 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 done to avoid introducing bian in our analysis. Because this is done at random we are not explicitly choosing our samples based on some criteria, we simply select at random odds. This also allows our samples to be representative of the population which we are trying to study.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
    - The stakeholders for this project are the management team who are looking for data oriented solutions towards retaining their employees and investing in those who are likely to stay.
- What are you trying to solve or accomplish?
    - I'm trying to find the characteristics of employees who stay with the company for a long time, as this can inform the management team on what type of employee they should invest time and training on.
- What are your initial observations when you explore the data?
    - The distributions of several employee statistics are very similar across all departments, which means that all departments share common characteristics regarding their employees. Such as number of projects, years with company, satisfaction levels, evaluations, history of accidents.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
    - Initially I was thinking of using ANOVA to test for statistical significance between the different values of satisfaction across departments but after seeing the distribution across all departments I opted not to because the distributions were bimodal, and highly irregular which meant they could not be treated as normal distributions. Sampling could've been done however, the all department shared the common bimodality and shape, which suggested they were not significantly different from one another
- Do you have any ethical considerations at this stage?

    - We need to ensure we include all departments in our analysis such that we are training the model to consider all these types of employees. This could potentially avoid discriminations in the classification or miss-classification in the performance of the model. Which could altogether affect how management treats that department.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
    - I'm trying to identify the features that have the highest relevance towards distinguishing employees who tend to leave and those who tend to stay.
- What resources do you find yourself using as you complete this stage?

  - I used the XGBooster notebook and the cross-validated random forest notebook from this course to orient myself in the process of building these models.
- Is my data reliable?
  - The data come from the company itself so such statistics are unlikely to be fabricated or a misrepresentation of the workforce present at the company.
- Do you have any additional ethical considerations in this stage?
  - We want to ensure our training data is inclusive and consider all types of employees present at the company. Since a machine learning model is as good as the data used to train it, it is important that we give representation to all employees.
- What metric should I use to evaluate success of my business objective? Why?
  - The best metrics that can be used to evaluate the success of the model are (accuracy, recall, precision, f1) and transparency of the model. At the end of the day even if we have an excellent model that can predict the turnover rates of employees, we still need to know what factors determine this, so we need a model that tells us how each feature related to the target variable.

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?
    - We believe that the features provided by the employer should suffice because we want features that the employer has direct control of such that they can manipulate it according to our suggestions. A feature that an employer does not have direct control over is employee satisfaction, the employer cannot automatically change this, there could be multiple factors that contribute to this statistic, some of which might not even be in our dataset, so it's best to have tangible features that can be changed according to our findings.

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?
    - We see that the years with a company have multiple outliers which need to be dealt with, as some models such as decision trees are very sensitive to this type of data.
- 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.?
    - There is no additional data outside what was provided to me that I need to add. The features provided are specific enough such that we can pinpoint areas of interest towards employee retention.
- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
    - Histograms and violin plots could come in handy to better understand the distribution of the variables and their relationship across different categories such as salary level and department. Additionally, heat maps for correlation and confusion matrices could come in useful as well. And bar plot for feature importance/relevance.

The Power of Statistics

- Why are descriptive statistics useful?

  - Correlation coefficient was one of the most important statistics for our analysis as well as IQR for detecting outliers in the cases of skewed and non-Gaussian distributions.
- What is the difference between the null hypothesis and the alternative hypothesis?
  - In the premise of studying satisfaction levels across different departments, the null hypothesis would be: “There is no significant difference between the satisfaction level of any department with the whole set of departments.”
  - The alternative hypothesis would simply be the converse of this statement.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
  - The purpose of EDA before performing any sort of regression is to ensure data is of high quality and without any bias which could affect the results produced by the regression. For example variables with strong multicollinearity do not satisfy the assumption of observational dependence which many regression models rely on.
- Do you have any ethical considerations at this stage?
  - We want to ensure that removing data during the EDA process does not introduce any bias or remove a population from the dataset, as this could lead to misclassification on that specific population.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
  - So far the analysis indicates that sales has a higher proportion of losing employees than any other department, and we also see that sales has the highest proportion of low income workers, however, this alone is not enough evidence to state this caused the high turnover ratio. Since the correlation coefficients are not that high with the target variable.
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
  - The EDA process ensured that the data followed all the assumptions of the model.
- Why did you select the X variables you did?
  - I picked them based on their correlation coefficients to the target variable and between themselves.
- What has the EDA told you?

  - There were to variables that were moderately correlated, so I created a new feature that represented the frequency of projects that employee had participated in the last month.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
- How many vendors, organizations or groupings are included in this total data?

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?
- 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)?

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
- What conclusion can be drawn from the hypothesis test?

Regression Analysis: Simplify Complex Data Relationships

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

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?
- Which independent variables did you choose for the model, and why?
- How well does your model fit the data? (What is my model's validation score?)
- Can you improve it? Is there anything you would change about the model?
- Do you have any ethical considerations in this stage?

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?
    - We observe that employee satisfaction, time spent with company, and hours per project are really good indicators for employee retention. The third feature is the most interesting one because this is something the company can directly manipulate by assigning more projects to a specific dpt. We see that the tech department has the best metric for this feature, so we recommend investing more resources in this dpt. as they are the ones most likely to continue with the company.
    - We also see that the sales department has more hours but very little projects to show for their hours and their correlation coefficient with the target variable indicate that this dpt. Is prone to employees that leave.
- What data initially presents as containing anomalies?
    - The main anomaly seen in this data was the amount of duplicated rows, which was over 300 rows. After inspecting the duplicated rows we saw that they were all identical and had to be dropped, it might've been a meta-data error
- What additional types of data could strengthen this dataset?
    - The results of our model indicate that the turnover rates are not related to financial reasons, they are more related to overall activity in the company. So more active members who collaborate in more projects are likely to stay. So it would be interesting to see the extra hours member put into their work such as working on weekends or holidays, because, these type of habits would suggest the member is very active in their work.

Go Beyond the Numbers: Translate Data into Insights

- What key insights emerged from your EDA and visualizations(s)?
    - The correlation matrix between the dpts. And the engineered feature hours per project provided a visualization on how each department is more active than others, as well as the direction of the correlation which is the most important aspect.
    - The bar plots also indicate the feature importance for predicting the longevity of employees and compares them directly to one another.
- What business recommendations do you propose based on the visualization(s) built?

  - Same recommendation as previously mentioned, tech department is the safest bet to invest time and resources into.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
  - We see that projects are a big factor for determining activity within the company, it would be interesting to see what other features we could engineer with time or projects separately to see any tendencies.
- How might you share these visualizations with different audiences?
  - For more technical audiences I would show the confusion matrix, table of results and the hyperparameters explored. With less technical audiences I would focus on the end recommendations.

Regression Analysis: Simplify Complex Data Relationships

- To interpret model results, why is it important to interpret the beta coefficients?
  - Since I didn't perform a logistic regression this doesn't apply however, beta coefficients are useful to quantify the relationship between a variable and a target variable.
- What potential recommendations would you make to your manager/company?
  - Same as before, tech department has higher chances for employees to stay, while the sales department has lowest chances to stay.
- Do you think your model could be improved? Why or why not? How?
  - Model could be improved by looking at how much revenue each department generates for the company. That way we can do a sort of optimization where we want to maximize revenue but minimize the risk of our investment into that department.
- What key insights emerged from your model(s)?
  - Salary does not play a significant role in the longevity of an employee, instead employee activity and engagement with company projects does.

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?
  - The random forest classifier outperformed the XGBoosted in all four evaluation metrics (Accuracy, Recall, Precision, F1) which meant picking a champion model was rather easy. We also noticed that the random forest classifier performed better in the test set than in the validation sets.
- What are the criteria for model selection?

  - As previously mentioned the 4-evaluation metrics and the transparency of the model in order to avoid having a black box.

- Does my model make sense? Are my final results acceptable?
  - The final results of my model make logical sense and they can be easily explained through reason. It was shocking to find that low salary as the 5th most relevant feature for employee retention.
- Were there any features that were not important at all? What if you take them out?
  - No feature was ignored in our analysis as we saw they all had comparable correlation coefficients; they all seemed like good predictors for the target variable. With the exception of satisfaction and and work accident
- Given what you know about the data and the models you were using, what other questions could you address for the team?
  - Because we used a random forest the feature importance represents the purity score, which means that feature does a really good job at segregating initial data into pure categories, which is the optimal approach for decision trees i.e random forests.
- What resources do you find yourself using as you complete this stage?
  - I used all the notebooks for XGBoosters, Random forest, tuning hyperparameters and cross validating models to avoid overfitting and better management of randomness in data.
- Is my model ethical?
  - The model is ethical as it considers all departments, and all types of employees such that no one single type of population is excluded from training the model. In addition the model was also trained for randomness, such that it can include odd points into the correct classification.
- When my model makes a mistake, what is happening? How does that translate to my use case?
  - If the model makes a mistake that means we are not qualifying a hard working employee to have longevity with our company and would therefore, lead to no efforts be put in place to maintain this employee. Which would just reinforce the employee to leave the company anyways.