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 project proposal

Overview

Salifort Motors wants to take some initiatives to improve employee satisfaction levels at the company. The purpose of this project is to generate a decision tree to predict if an employee will leave the company and determine the most important factors that determine employee satisfaction.

Milestones	Tasks	PACE stages
1	Establish structure to project workflow (PACE)	Plan
1a	Write a project proposal	Plan
2	Familiarize and understand the available data	Analyze
3	Data exploration and cleaning (EDA)	Plan & Analyze
4	Build a machine learning model	Construct
5	Build an executive summary	Construct
6	Communicate final insights with stakeholders	Execute

Data Project Questions & Considerations

PACE: Plan Stage

- **What am I trying to solve?**

  The purpose of this project is to predict whether an employee is likely to leave the company. Additionally, it is important to determine the most important factors for the position.

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

  The best model to predict this situation is a Decision Tree. This is because a Decision Tree model will have good metrics at predicting the outcome variable. In addition, this model does not require a booster, because the use of a booster limits the understanding of the impact of independent variables on the objective categorization.

- **Is my data reliable?**

  Yes, this is because the model covers relevant variables which seem to be related to the outcome variable. Furthermore, a survey is a proper collection method for this kind of project.

- **What data do I have/can I get?**

  I have relevant features such as the dependent variable (whether the employee left the company) and several important variables related to the employee's satisfaction, performance, salary, among others.

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

  Recall would be the most important metric because it focuses on the proportion of positive values predicted, so the main focus in the project are employees who are likely to leave the company.

Data Project Questions & Considerations

PACE: Analyze Stage

- **Why did you select the X variables you did?**

  Because they seem to have a strong relationship with the outcome variable, so they can improve the accuracy of the model.

- **What has the EDA told you?**

  1. Every employee who worked on 7 projects left the company.

  2. There is a significant difference between those who stayed or left: When the number of projects increased, those employees who stayed maintained a constant number of hours worked. On the other hand, those who left worked more hours as the number of projects increased.

  3. Employees who worked significatively less time than those who stayed finally left.

  4. The optimal median of worked hours per month is approximately 190-200 hours.

  5. The optimal number of projects per worker are between 3 and 4. While the worst are 2, 6 and 7.

  6. There is a huge difference between the employees who worked on 2 projects compared to those who worked on 3 projects.

  7. People who left the company were divided into 3 clusters.

  7.1. People who worked less than the average worked hours and had a satisfaction of 0.4.

  7.2. This satisfaction can be caused by a pressure from directives or maybe lower salaries due to their schedule.

  7.3. People who worked more than 240 hours and showed a minimum satisfaction. This can be due to an over workload.

  8. Satisfied people who worked between 230-270 hours and were satisfied (0.8). This cluster of people is very strange because of their good performance and good satisfaction.

  9. Those who left the company where divided into 2 clusters:

9.1. Employees who worked less time than the normal and had a bad score (lower than 0.6). This can be due to the fact that they could be fired because of their low performance by level and worked hours.

10. Employees who worked more than 230 hours and scored very well (higher than 0.8). In part this group would be composed of 2 subgroups: those who left the company because overwork (this can be correlated to a lower satisfaction level like the cluster of the last graph), and another group of people who worked more than the normal and had a higher performance (this group can be related to the rare cluster of the last graph, which was characterized also by higher satisfaction levels).

11. Those who worked more than approx. 280 hours per month were not promoted, which is a strong factor that could explain the low satisfaction of the cluster with more hours worked. This is important to evaluate because it is necessary to understand why they were not promoted, and the most important fact, those who worked more showed a higher performance, which means that their overwork does not mean a lack of proactive work. This is very important to evaluate because this can be a key factor.

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

  In this project I focused on visualization cleaning because it is an interesting way to analyze and it is good to evaluate key features of classification, something that helps in the case of a decision tree model.

- **Do you have any ethical considerations at this stage?**

  I am starting to question the methods used to evaluate the metrics of the employees. This is because there are variables with subjective values such as the salary, satisfaction, among others, also, there are variables that can be hard to evaluate such as the workload.

Data Project Questions & Considerations

PACE: Construct Stage

- Do you notice anything odd?

There is a group of employees who had good metrics (including satisfaction) and left the company.

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

I decided to choose all of the variables except satisfaction level because of its huge importance in the model but subjectivity.

- How well does your model fit the data?

Perfectly, it has pretty good metrics and a significant recall score.

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

Principally packages related to decision trees of Sklearn.

- Do you have any ethical considerations at this stage?

It seems that the principal reason for leaving the company is overworking. So the company can be deficient at the moment of showing the basis of their contracts.

Data Project Questions & Considerations

PACE: Execute Stage

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

The most important aspects to improve the employees' retention are the satisfaction level (which is subjective, so it is useful only to identify the situation of the employee), number of projects, score in the last evaluation, tenure, and whether the employee is overworked.

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

It is important to evaluate the workload of the employees. This is because most of the features related to staff exit are related to overworking and number of projects. In the case of being overworked it is seen that surprisingly most of the staff is overworked, working more than 180 hours per month. In the case of the number of projects is essential to consider that to maintain the employees the number of hours worked must be constant when the number of projects increase; also, the number of projects should be closer to 3 or 4, and the number of projects must avoid 7 projects or more (in that case all the employees left the company). In the case of tenure, it is appreciated that there is a crucial error with employees with large ternures. This is because they were not promoted or their salary did not grow despite showing a great performance, something rare considering their loyalty and performance in the company. Finally, the results in the evaluations are important to recognize that in some cases there could be a problem with the capacitation of the staff, it is important to create incentives to make them improve their skills.

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

It is important to investigate the cluster of employees who had perfect stats (including satisfaction) and left the company. Also it is important to evaluate why people with more experience in the company have not been promoted despite their loyalty and good performance. Also it is important to evaluate the case of overworking, which is the main cause of the poor staff retention.

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

Of course, I consider that this model is propense to overfitting due to the nature of decision trees, so a good approach could be the development of a random forest model in rounds. In the case of the present model it was considered that the Recall metric was working pretty well and there is no need to create a model that consumes more time and effort. Furthermore, the model can be also improved by analyzing the nature of the variables more deeply with the help of an expert in the subject. Another thing to consider is that in this case the evaluation score cannot be totally controlled by the managers, so it wouldn't be a good measure for the model, so it can be deleted to create a new round. Finally, a k-means model would be a perfect way to evaluate the case of the clusters of employees who left the company and also had good scores.