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 analyse 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

To build a Classification Machine Learning Model for Salifort Motors, that can predict when an employee will leave the company.

Milestones	Tasks	PACE stages
1	• Establish structure for the project workflow (PACE) • Write a project proposal	Plan
2	• Compile summary information about the data • Explore the data	Plan Analyze
3	• Data exploration and cleaning • Compute descriptive statistics • Visualizations	Plan Analyze
4	• Data Preparation • Splitting the data into Train, Validation and Test Sets • Build a Logistic Regression Model • Build Tree-Based Machine Learning Models • Conduct Hyperparameter Tuning and Cross-Validation	Construct
5	• Compute GridSearch Evaluation Metrics • Compute Model Prediction Performance Evaluation Metrics • Champion Model Identification • Feature Engineering	Construct

	• **Build Feature Engineered Tree-Based Models**	
6	• **Compute GridSearch Evaluation Metrics** • **Compute Model Prediction Performance Evaluation Metrics** • **Champion Model Selection** • **Compute Champion Model Evaluation Metrics** • **Plot Feature importance**	**Construct**
7	• **Recommendations** • **Communicate final insights with stakeholders**	**Execute**

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 senior leadership team at Salifort Motors.
- **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 which employees are at a high risk of leaving Salifort Motors so that leaders can intervene early with targeted actions instead of reacting after resignations. The expected impact is lower unwanted employee turnover, reduced hiring and training costs, better retention of top performers and more stable teams, thereby directly supporting productivity, service quality and long-term business performance.
- **What questions need to be asked or answered?** → What are the most predictive features leading to an employee leaving Saliforts? What is the impact of misclassifying a leaver as an employee who stayed?
- **What resources are required to complete this project?** → pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost and pickle.
- **What are the deliverables that will need to be created over the course of this project?** → EDA of the data as well as visualizations, Modelling and Model Evaluation Metrics from both the GridSearch and the Model predictions.

Get Started with Python

- **How can you best prepare to understand and organize the provided information?** → I can best prepare to understand and organize with initial exploratory data analysis (EDA).
- **What follow-along and self-review codebooks will help you perform this work?** → The Jupyter Notebooks from the certificate that cover both EDA and Modelling, such as the Course 6 Jupyter Notebook materials.

- **What are a couple additional activities a resourceful learner would perform before starting to code?** → Writing both a project proposal and PACE Strategy document and writing and understanding the project scenario, data dictionary and objectives. Another activity would be to look at the documentation of the libraries or functions the learner plans to use in the project, as to see if any additional code can be added.

Go Beyond the Numbers: Translate Data into Insights

- **What are the data columns and variables and which ones are most relevant to your deliverable?** → They are: *satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, Work_accident, left, promotion_last_5years, Department, salary*. The most relevant one is *left* as it's the target variable for the project as it is a categorical variable that states whether or not an employee left the company.
- **What units are your variables in?** → The dataset uses the following units/scales: *satisfaction_level* and *last_evaluation* are unitless scores on 0–1, *number_project* is a count, *average_monthly_hours* is measured in hours per month, *time_spend_company* is in years, *work_accident, left,* and *promotion_last_5years* are binary indicators (0/1), *department* is a categorical field (e.g., sales, technical) and *salary* is a categorical band (low/medium/high).
- **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 presumptions are that satisfaction, number of projects, working hours and promotion will be the most predictive features, with promotion also acting as a protective feature against employee's leaving.
- **Is there any missing or incomplete data?** → There is no missing data
- **Are all pieces of this dataset in the same format?** → No as there is non-numerical data such as *salary* and *department*.
- **Which EDA practices will be required to begin this project?** → Discovering, Cleaning and Validating.

The Power of Statistics

- **What is the main purpose of this project?** → To predict employee turnover for Salifort Motors.
- **What is your research question for this project?** → How to increase employee retention?

- **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 important as it gives every member of the population a chance of selection which in turn reduces bias and makes the sample a representative sample of the population. A representative sample accurately reflects the characteristics of the population, and Models based on representative samples are much more likely to make fair and unbiased decisions. Sampling bias is when a sample is not representative of the population as a whole. So when you use non-probability sampling methods such as Convenience Sampling and because these are based on convenience to the researcher and not a broader sample of the population, these samples often show undercoverage bias. This is when some members of a population are inadequately represented in the sample.

Regression Analysis: Simplify Complex Data Relationships

- **Who are your stakeholders for this project?** → Salifort Motors Senior Leadership team.
- **What are you trying to solve or accomplish?** → To analyse employee survey data and to predict which employees are likely to leave the company. This involves identifying key factors driving employee turnover, building an accurate predictive model and providing data-driven recommendations to improve retention and reduce turnover costs. Overall, the project aims to help Salifort Motors to develop a sustainable strategy for employee engagement and retention.
- **What are your initial observations when you explore the data?** → My initial observations are that there are some categorical variables that will need to be encoded, some columns need to be renamed so that they are standardised, there are no missing values, duplicates were removed and there are outliers in the data. There may also be a class imbalance in the *left* feature as fewer people may have stayed and also features such as *promotion_last_5years* and *Work_accident* could act as protective features against leaving.
- **What resources do you find yourself using as you complete this stage? (Make sure to include the links.)** →
    - Pandas
    - NumPy
    - Matplotlib.pyplot
    - Seaborn

- Scikit-learn.model_selection
- Scikit-learn.metrics
- Scikit-learn.linear_model
- Scikit-learn.ensemble
- Scikit-learn.tree
- Xgboost
- pickle

- **Do you have any ethical considerations in this stage?** → Yes I have several ethical considerations at this stage such as the data privacy of the employee data which must be handled securely and anonymised to protect their identities. Another consideration is Bias as the model should not unfairly discriminate based on non-performance related features such as *Department* or *salary*. There should also be transparency with the Model results by them being interpretable and used responsibly to support the employees and not to penalise them.

The Nuts and Bolts of Machine Learning

- **What am I trying to solve?** → To classify employee turnover at Salifort Motors.
- **What resources do you find yourself using as you complete this stage?** → Pandas, NumPy, Matplotlib.pyplot, Seaborn, Scikit-learn.model_selection/metrics/tree/ensemble/linear_model, xgboost and pickle.
- **Is my data reliable?** → Yes as it is 1st party survey data from Human Resources where they surveyed a sample of employees.
- **Do you have any additional ethical considerations in this stage?** → Yes there are several ethical considerations at this stage. Privacy: employee data must be handled securely and anonymized to protect identities. Bias: the Model should not unfairly discriminate on non-performance factors (*department* or *salary*). Transparency and Use: results should be interpretable and applied responsibly to support employees such as guiding coaching, workload and development rather than to penalize them.
- **What data do I need/would I like to see in a perfect world to answer this question?** → More of the employees who left so the classes can be more balanced, this will help with Model performance as more of the features of left employees can be studied.
- **What data do I have/can I get?** → I have the Human Resources Survey data.

- **What metric should I use to evaluate success of my business objective? Why?** → The main evaluation metric to focus on for the success of my business objective is Recall. This is because it is more costly for Salifort when there is a False Negative (missing a leaver) and so by focusing on Recall it maximizes the fraction of leavers that are correctly identified.

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 as I have been able to explore the relationships between the variables and the target variable *left* and have also been able to visualize them and gain insights from these data visualizations. Based on this analysis of the variables, I will be able to understand the data better when modelling and also know what features to not include in my Model such as *department*.

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?** → Checking and removing null and duplicate data, checking for outliers and decide whether or not to remove them based on what modelling technique is used and the Model's assumptions. The class balance of the target variable needs to be checked so that it can be rebalanced if there is a severe class imbalance. Finally, data visualizations such as box plots, bar charts, scatterplots and heatmaps to be created displaying the relationships between the variables as this will give us useful insights and understanding of the 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.?** → No, as no more data needs to be added via Joining as there is no available data to be joined, otherwise it would be a yes and then would be joined. The only type of EDA Structuring that needs to be done is when using Boolean Masking/filtering when using the IQR method for the outliers and when checking the class balance of the target variable.
- **What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?** → Box plots will be best for the numerical data points as this can show the distribution and the measures of position for the data points. For categorical data, grouped bar charts with the count on the y-axis and the Categorical variable (*salary*, *promotion, department*) on the x-axis grouped by the target variable *left*. For example, this shows the count of how many employees with a low salary left the company as its all displayed in one plot. Scatterplots to show how key drivers like satisfaction and hours jointly relate to leaving. Finally, a heatmap of all the numerical data which shows us the correlations between the variables.

The Power of Statistics

- **Why are descriptive statistics useful?** → Descriptive statistics are useful because they give us a quick insight into the data, through measures of central tendency, dispersion and position. This allows us to spot issues in the data really quickly and give us a better understanding of the dataset as a whole. This in turn helps guide the modelling process and choices (feature engineering, encoding) as well as clear communication with stakeholders.
- **What is the difference between the null hypothesis and the alternative hypothesis?** → The null hypothesis is the default claim of no effect or no statistical significance, while the alternative hypothesis is opposite claim of there being an effect or there being a statistical significance.

Regression Analysis: Simplify Complex Data Relationships

- **What are some purposes of EDA before constructing a multiple linear regression model?** → Some of the purposes are to check for the Multiple Linear Regression Model assumptions, some of which is done before constructing a Linear Regression Model such as Linearity with a scatterplot matrix produced by seaborn's .pairplot() function. Also for the OLS function from statsmodels.formula.api, in the EDA stage the preparation of the data for the OLS function can be done and then assigned as "ols_data" to be used in the function.
- **Do you have any ethical considerations in this stage?** → Yes I have several ethical considerations at this stage such as Data Privacy with employee data being handled securely, access-controlled and anonymised in order to protect the employees identities, with clear purpose limitation and minimal collection. I will also monitor and mitigate for bias by ensuring the Model does not unfairly disadvantage groups based on non-performance features such as `department` or `salary`, this is done using fair sampling, class weights and checking performance by subgroup. Transparency also matters and so the Model should be interpretable and its outputs communicated clearly so they're used to support employees and not to penalise them. Finally, I will also plan for ongoing monitoring in order to keep the system fair, private and accountable over time.

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 solve employee turnover rate for Salifort Motors and yes the plan of using a classification Model still works. No as this will only be necessary if one of the assumptions of the Model is violated.

- **Does the data break the assumptions of the model? Is that ok, or unacceptable?** → Some of the assumptions are partially broken for the Logistic Regression Model such as non-linear relationships between features and turnover, which might lead to weaker performance and make it less suitable as the primary Model. For the Tree-Based Models, these issues are not a problem as they do not rely on linearity or normality assumptions and so by using them as the main modelling approach is acceptable and appropriate for this dataset.
- **Why did you select the X variables you did?** → I selected them as they displayed having a great relationship with the target variable *left* of the employees who left Saliforts.
- **What are some purposes of EDA before constructing a model?** → Some of the purposes are to check to see if the data fits the Models assumptions, to better understand the features and how they interact with the target variable, to aid in feature selection, transformation and extraction.
- **What has the EDA told you?** → The EDA has told me that for the target variable, there are 10,000 (83%) employees who stayed and 1,991 (17%) employees who left. There are outliers in the variable *tenure* as well as 3008 duplicate rows in the whole dataset that were removed and no null values. The EDA also showed me the relationship between the target variable and *satisfaction_level, last_evaluation, tenure, number_project, average_monthly_hours, salary, promotion_last_5_years* and *department*. These visualizations lead to many insights from each and also showed that *satisfaction_level*, workload, burnout as well as lack of recognition can lead to employee turnover. The final visualization of a heatmap showed the strong negative correlation between the target variable and *satisfaction_level* indicating it is a major factor in employees leaving.
- **What resources do you find yourself using as you complete this stage?** → pandas, NumPy, Matplotlib.pyplot, Seaborn.
- **Do you have any ethical considerations in this stage?** → At this stage my priorities are ethical such as Data Privacy ensuring secure and access-controlled handling of employee data with anonymisation, minimal collection and a clearly defined purpose. Bias mitigation helps to ensure the Model doesn't disadvantage groups on non-performance factors such as *department* or *salary* by applying fair sampling/class weights and auditing subgroup performance. Transparency helps keep the Model interpretable and to be able to communicate the results plainly so that they're used to support and not penalise employees. Finally, ongoing oversight of the Model such as continuously monitoring performance, fairness and privacy in order to keep the system accountable over time.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- **Do any data variables averages look unusual?** → N/A
- **How many vendors, organizations or groupings are included in this total data?** → In the data the only organization that is included is Salifort Motors as it is their Human Resources Team that collected the data via surveys.

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?** → In order to complete the project goals at this stage, the data visualizations that would be needed are Confusion Matrix's and a Feature importance plot. The Machine learning algorithm needed will be a Decision Tree, Random Forest and XGBoost Classifiers. These will be hyperparameter tuned via *GridSearchCV* and its evaluation metric results from the GridSearch will be saved in a dataframe as to compare Model performance. Another data output is the Models predictions performance via evaluation metrics on the validation sets in order to compare Model predictive performance and choose a champion Model.
- **What processes need to be performed in order to build the necessary data visualizations?** → The data needs to be assigned into their X and Y variables, and then split into Training, Validation and Testing datasets. The champion Model then needs to be instantiated and then fitted to the training data using the *GridSearchCV* function, in order to get the best estimator. Afterwards, Model predictions need to be made on the validation X data for Y and these need to be compared with the actual validation Y data. Finally, these results need to be inputted into the 2 Confusion Matrix functions (*confusion_matrix*, *ConfusionMatrixDisplay*) and then plotted to produce the Confusion Matrices. The same process is repeated for the test X and Y data with Model predictions made and then plotted on a Confusion Matrix. For the feature importance plot, it can be easily produced from the champion Model, via the attribute *.feature_importances_* and plotting it with *sns.barplot*.

- **Which variables are most applicable for the visualizations in this data project?** → In this case, all the variables as we are visualizing the Model predictions and Model feature importance.
- **Going back to the Plan stage, how do you plan to deal with the missing data (if any)?** → I plan to investigate first to see if the missing data should be removed and if after careful consideration it needs to be removed, I will remove the missing data in order to clean the dataset.

The Power of Statistics

- **How did you formulate your null hypothesis and alternative hypothesis?** → N/A
- **What conclusion can be drawn from the hypothesis test?** → N/A

Regression Analysis: Simplify Complex Data Relationships

- **Do you notice anything odd?** → Yes, in the first round of Modelling, all three of the Tree-Based Models achieved extremely high scores across Accuracy, Recall and ROC AUC on both training and validation sets. This is unusual and therefore raised concerns about potential data leakage or over-reliance on a few overly powerful features rather than learning patterns that would generalize.
- **Can you improve it? Is there anything you would change about the model?** → To address this, I redesigned the features via feature engineering I dropped satisfaction, I reduced the reliance on leakage-prone signals and created more realistic, business-friendly variables such as the `*overworked*` flag and re-ran the Models. This feature-engineered setup produced still strong but more credible performance, with the feature-engineered Random Forest Model emerging as a stable, interpretable champion Model that better reflects how HR could act on the results.

The Nuts and Bolts of Machine Learning

- **Is there a problem? Can it be fixed? If so, how?** → Yes, the first round of Tree-Based Models looked almost “too-good” therefore suggesting possible leakage or over-reliance on raw variables like *average_monthly_hours* rather than robust patterns. We addressed this by engineering more realistic features, re-running Cross-Validated Models and validating on a held-out validation set, this in turn produced strong but more credible results and reduced the risk of overfitting.

- **Which independent variables did you choose for the model, and why?** → We included evaluation score, tenure, number of projects, overworked, promotion history, salary level, department and work accidents because they directly capture workload, performance, reward and experience which are the core drivers of an employee's decision to stay or leave and are available and actionable for HR levers.
- **How well does your model fit the data? (What is my model's validation score?)** → The feature-engineered Random Forest champion Model, achieves a high and balanced performance. On the validation set it reaches roughly the mid-90s Accuracy with strong Recall and ROC AUC in the low-to-mid 0.9s and on the unseen test set it maintains about 97% Accuracy, 0.98/0.91 Precision, 0.98/0.88 Recall (stay/leave), F1-Score around 0.97/0.90 and ROC AUC around 0.96, thereby indicating excellent generalisation to new employees.
- **Can you improve it? Is there anything you would change about the model?** → We can further improve the Model by tuning the decision threshold to favour catching more true leavers, monitoring performance over time for drift, adding richer features (e.g. internal mobility, training) to capture context and routinely checking fairness and explainability so HR, Legal and Leadership can trust and act on the predictions.
- **Do you have any ethical considerations in this stage?** → The ethical considerations to be thought of in this stage would be the aforementioned considerations mentioned before as well as monitoring the Model after deployment. This would entail tracking the Model performance by segment in order to keep the system fair and effective.

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 would recommend to the senior leadership team to investigate any more additional factors that could lead to an employee leaving such as social factors like being treated differently based on sex, race or gender or bullying. This will give the senior leadership team insights into the workplace culture that employees may be facing and can lead to them implementing changes and training for employees about Salifort's values.
- **What data initially presents as containing anomalies?** → The data that initially presented as containing anomalies was the feature *tenure*. *tenure* was visualized on a boxplot visualization in order to see its distribution and it was evident there are outliers in the upper bound. This was further confirmed by doing the Interquartile range method for finding outliers, with *tenure* having 824 outliers in its upper bound.
- **What additional types of data could strengthen this dataset?** → Richer data on the employee journey would make the Model much stronger, for example, longitudinal trends in satisfaction and performance, detailed promotion and role-change history, manager and team information, absenteeism and overtime patterns, training/engagement programs participation, internal mobility and applications and perhaps external factors like local labour market conditions or compensation benchmarks to better capture why and when the high-value employees decide to leave.

Go Beyond the Numbers: Translate Data into Insights

- **What key insights emerged from your EDA and visualizations(s)?** → The employees who left tended to have lower satisfaction, higher monthly hours and either very few or many projects. They are often mid-tenure, slightly higher-rated thereby suggesting overwork or lack of recognition and less likely to have been promoted in the last five years. Lower/medium salary bands show higher leaving rates, while department effects are modest with it being slightly higher in *sales, technical* and *support*. The heatmap confirms satisfaction is the strongest driver via a negative correlation with *left* with a mild positive link for *tenure*.

- **What business recommendations do you propose based on the visualization(s) built?** → The confusion matrix shows the Model rarely flags stayers incorrectly and correctly identifies most leavers, while the feature importance highlights the number of projects, tenure, last evaluation and overworked hours as key drivers. Based on this, Salifort should monitor the employees with high project loads, long tenure (3-5+ years), strong evaluations and overworked status as priority retention cases. They can also proactively adjust workload and staffing, ensuring high performers are not chronically overloaded and create clearer growth and promotion paths for mid-tenure high performers. Finally, Salifort should embed the Model into the HR processes as an early warning tool to trigger supportive interventions and not punitive actions.

- **Given what you know about the data and the visualizations you were using, what other questions could you research for the team?** → I would investigate which specific combinations of risk factors (e.g. high projects + overworked + mid-tenure + high evaluation) lead to the highest predicted leave risk, how risk varies by manager and department, what interventions (promotion, workload change, pay adjustment) most effectively reduces that risk over time and whether the Model behaves fairly across gender, role and location segments.

- **How might you share these visualizations with different audiences?** → For executives, I'd show a simple confusion matrix summary and a top-5 feature importance chart with plain language takeaways and impact estimates. For HR and People Managers, I'd provide a slightly richer dashboard combing these plots with drill-downs to team/employee segments and for the data and analytics stakeholders, I'd share the full confusion matrix, feature importance and methodology so they can validate performance, bias and stability.

The Power of Statistics

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

- **What business recommendations do you propose based on your results?** → Use the model to proactively flag high-risk employees and trigger targeted retention actions rather than reacting after resignations. Focus on workload management (limit sustained overtime and project overload), especially for strong performers whose high evaluations plus heavy hours signal flight risk. Strengthen career paths and promotion opportunities for mid-tenure employees, who show elevated risk if progression stalls. Review compensation for critical roles to ensure pay is competitive relative to workload and performance. Finally, embed regular monitoring and fairness checks so the model supports people decisions transparently without driving biased or punitive actions.

Regression Analysis: Simplify Complex Data Relationships

- **To interpret model results, why is it important to interpret the beta coefficients?** → It is important to interpret the beta coefficients when interpreting the Model results because it tells you how each predictor is associated with the target variable in direction and size while holding the other variables constant. This in turn lets you translate the model outputs into business terms such as "a one-unit drop in satisfaction changes the log-odds of leaving by …", and it also lets you compare the relative influence of the features. Examining beta coefficients with Confidence Intervals and P-Values also supports hypothesis tests and helps decide which leavers are meaningful to act on, which to drop and where interactions or nonlinearity may be needed. In other words, beta coefficients turn a fitted model into interpretable, actionable insights rather than just a prediction score.

- **What potential recommendations would you make to your manager/company?** → To use the champion Model as an early-warning system to flag high-risk employees and trigger supportive interventions and not punitive actions. Prioritise improvements in workload management (limit chronic overtime and project overload), strengthen promotion and career paths for mid-tenure staff, protect and reward high performers who show burnout risk and review compensation in key roles as to stay competitive. Embed clear governance via documented rules for how risk scores are used, regular fairness checks and transparent communication with employees.

- **Do you think your model could be improved? Why or why not? How?** → Yes, while the performance is strong and more realistic after the feature engineering, we can further improve by tuning the classification threshold around business costs, exploring cost-sensitive learning as to better prioritise catching true leavers, calibrating probabilities and adding richer features to capture context that the current dataset misses. Ongoing monitoring for drift and periodic retraining are also needed to keep predictions reliable and fair over time.

- **What business recommendations do you propose based on the models built?** → Based on the models, Salifort should use predictions to identify high-performing employees with heavy workloads and intervene early with workload rebalancing, recognition, and dedicated career conversations to reduce burnout-related exits. The strong role of tenure and performance suggests investing in clear career paths and timely promotions for mid-tenure, high-performing staff who have not advanced. The “overworked” signal should be turned into concrete workload guardrails, such as limits on sustained overtime and project load, with manager alerts. Where the model indicates elevated risk in specific roles or lower salary bands, HR should review pay competitiveness and design targeted retention actions rather than blanket increases. Finally, model risk scores should be embedded into an HR dashboard—with strict governance and bias monitoring—so leaders receive interpretable, ethical, and regularly validated insights to guide supportive retention decisions.

- **What key insights emerged from your model(s)?** → The Models consistently show that project load, tenure, last evaluation and being overworked are the strongest drivers of leaving, while department and salary level play a smaller but non-negligible role. High-performing employees with heavy workloads and no recent advancement, especially at mid-tenure, are at a elevated risk and the champion Model can accurately identify most stayers and a large share of leavers, thereby giving Salifort a practical and interpretable tool to target retention where it matters most.

- **Do you have any ethical considerations at this stage?** → The ethical considerations I have at this stage are privacy through the anonymisation of employee data, least-privilege access and documented purpose limitation. Another consideration is fairness through avoiding protected attributes, evaluate Precision and Recall by subgroup (salary band, tenure) and mitigate disparities (thresholds, weights). Transparency and use of the Model is another consideration as the Model's outputs should be used to support the employees and not to penalise them and so its important to communicate how the Model is going to be used. Finally, governance and monitoring considerations such as log decisions, HR/legal review and continuously monitoring performance and parity after deployment.

The Nuts and Bolts of Machine Learning

- **What key insights emerged from your model(s)?** → The Models confirm that high project load, long tenure without progression, strong performance combined with overwork and being

consistently "overworked" are key predictors of leaving, while salary and department play smaller roles. This therefore suggests that Salifort's challenge is retaining high-contributing employees who are stretched and not advancing.

- **What are the criteria for model selection?** → I prioritised high and stable Recall for employees who leave with strong overall ROC AUC and F1-Score consistently between Cross-Validation/Validation/Test results (to rule out leakage) and interpretability that is suitable for HR and Leadership. On this basis, the feature-engineered Random Forest was chosen as the champion Model.

- **Does my model make sense? Are my final results acceptable?** → Yes the champion Model achieves 0.97 Accuracy and strong Recall (0.88) for leavers on the test set, with similar scores on the validation and Cross-Validation, which in turn aligns with the patterns seen in the EDA (overwork, tenure, performance, projects). The results are both plausible and operationally acceptable for guiding targeted retention actions.

- **Were there any features that were not important at all? What if you take them out?** → Department dummies and *work_accident* show very low importance and dropping them has little effect on the performance and can simplify the Model slightly. Core drivers such as *number_project, tenure, last_evaulation* and *overworked* should be retained.

- **Given what you know about the data and the models you were using, what other questions could you address for the team?** → We could explore which specific teams/managers have concentrated risk and how risk changes over time such as pre-exit trajectories, whether interventions reduce predicted risk, how performance vs. burnout trade-offs and whether the Model behaves fairly across locations, genders or roles.

- **What resources do you find yourself using as you complete this stage?** → The resources I used are:

    - Pandas
    - NumPy
    - Matplotlib.pyplot
    - Seaborn - barplot
    - Scikit-learn – Random Forest Classifier
    - Scikit-learn - GridSearchCV
    - Scikit-learn – confusion_matrix
    - Scikit-learn – ConfusionMatrixDisplay
    - Scikit-learn – classification_report
    - Scikit-learn – roc_auc_score
    - Scikit-learn – precision_score
    - Scikit-learn – recall_score
    - Scikit-learn – f1_score
    - Scikit-learn – accuracy_score

- **Is my model ethical?** → The Model is ethical when used with safeguards as it relies on job-related features, avoids sensitive attributes, protects employees privacy and is intended to flag people for support and retention and not for punitive actions while ongoing fairness checks ensure that no group is systematically disadvantaged.

- **When my model makes a mistake, what is happening? How does that translate to my use case?**
  → A False Positive is an employee flagged as a high-risk who would have stayed anyway and in practice they may receive extra support or recognition which is generally low-risk. A False Negative is a leaver that the Model misses, this is much costlier as the company loses someone without intervention and so tuning emphasises Recall for leavers while keeping False Positives at a manageable level.