Lloyds Banking Group

Data Science

**Task 1: Data gathering and Exploratory Analysis**

Task overview

Welcome! In this task, you will focus on gathering and analysing data to understand customer behaviour, with the goal of uncovering key insights that will inform the development of a predictive model for customer churn.

**What you'll learn**

* Techniques for identifying and collecting relevant data from various sources
* Methods for performing exploratory data analysis (EDA) to uncover patterns and insights

## Best practices for cleaning and preparing data for machine learning models  What you'll do

* Identify and gather data from provided sources relevant to predicting customer churn
* Perform EDA to understand the data and identify key features
* Clean and preprocess the data to ensure it is ready for model building

As you step into the role of a data science graduate at Lloyds Banking Group, you're immediately thrust into a real-world scenario with significant implications for our business. The Data Science & Analytics team, under the leadership of Li, a seasoned senior data scientist, is currently grappling with a critical project: predicting customer churn to enhance retention strategies.

Li briefs you on the situation, "We've observed a worrying trend of customers, particularly young professionals and small business owners, leaving for competitors. This project aims to reverse that trend by identifying at-risk customers and implementing targeted interventions."

Your task is crucial. You'll begin by gathering and analysing data to understand the factors contributing to customer churn to uncover actionable insights that can inform strategic decisions. The pressure is on, as SmartBank, a key subsidiary, has reported a decline in retention rates, and there's mounting pressure from senior management to deliver solutions swiftly.

# Techniques for identifying and collecting data

#### **Understanding the data landscape**

The first step in data collection is to understand the landscape of available data. In a corporate environment, data can come from a variety of sources, including internal databases, customer relationship management (CRM) systems, financial records, web analytics, and external data sets. Knowing where data resides and how it can be accessed is crucial for gathering relevant information efficiently.

#### **Defining data requirements**

Clearly define what you need to know before you begin. To predict customer churn, identify key variables such as customer demographics, transaction history, customer service interactions, and usage patterns. This will help focus your efforts on collecting the most relevant and useful data for your analysis.

#### **Data collection methods**

Data can be collected through several methods:

* **Primary data collection:** This involves gathering data directly from the source. For example, conducting surveys and interviews, or direct observation. While primary data is specific and relevant, it can be time-consuming and costly to obtain.
* **Secondary data collection:** Involves using existing data collected for another purpose. This could include historical data, third-party data sets, or data obtained from public sources. Secondary data is often easier and quicker to access but may require more careful consideration to ensure its relevance and accuracy.
* **Automated data collection:** Using tools such as web scraping, application programming interfaces (APIs), or data integration platforms to automatically gather data from online sources or databases. This method is efficient for collecting large volumes of data and keeping it up-to-date.

#### **Evaluating data quality**

Once data is collected, it's essential to evaluate its quality. Consider the following aspects:

* **Accuracy:** Ensure that the data accurately reflects the real-world scenarios you are studying.
* **Completeness:** Check for missing values or incomplete records that might skew your analysis.
* **Consistency:** Ensure that data is consistent across different sources and time periods.
* **Timeliness:** Data should be current and relevant to the time frame of your analysis.

#### **Data integration and storage**

After collecting data from various sources, the next step is to integrate it into a cohesive data set. This may involve merging data sets, transforming data formats, or cleaning data to ensure uniformity. Once integrated, store the data securely, ensuring that it is organised and easily accessible for analysis.

By mastering these techniques, you'll be well-prepared to gather and utilise data effectively, setting the stage for a successful data analysis project.

# Methods for performing exploratory data analysis to uncover patterns and insights

Exploratory data analysis (EDA) is a crucial stage in the data science process. It allows you to understand the underlying configuration of your data and identify key patterns and relationships. EDA involves using statistical techniques and data visualisation to summarise the main characteristics of the data set.

#### **Descriptive statistics**

Begin your EDA by calculating descriptive statistics, which provide a summary of the basic features of your data. Key metrics include:

1. **Measures of Central Tendency: Mean, Median, and Mode**

* **Mean, median, and mode:** These measures of central tendency help you understand the typical value in your data set.

2. **Measures of Dispersion: Standard Deviation and Variance**

* **Standard deviation and variance:** These metrics indicate the spread or dispersion of your data, showing how much variation exists from the average.

3. **Measures of Range: Min, Max, and Range**

* **Min, max, and range:** These values provide insights into the bounds of your data, highlighting any potential outliers.

#### **Data visualisation**

Visualisation is a powerful tool in EDA, helping you to quickly grasp complex data distributions and relationships. Common visualisation techniques include:

* **Histograms and density plots:** Useful for understanding the distribution of a single variable, including its central tendency and spread.
* **Box plots:** Help identify the spread and outliers in your data by showing the quartiles and median.
* **Scatter plots:** Ideal for examining relationships between two continuous variables, helping you identify potential correlations or trends.
* **Bar charts and heatmaps:** Useful for categorical data, showing the frequency or proportion of categories and the relationships between categorical variables.

#### **Correlation analysis**

To uncover relationships between variables, perform a correlation analysis. This involves calculating correlation coefficients, such as Pearson’s or Spearman’s, which quantify the strength and direction of the relationship between variables. Understanding these correlations can help you identify key predictors of customer churn.

#### **Data profiling and anomaly detection**

Data profiling involves examining data for anomalies, missing values, or inconsistencies. This phase is essential for ensuring data quality before proceeding to model building. Look for patterns in missing data, which could indicate underlying data collection issues or important missing information.

#### **Hypothesis generation**

Based on the insights gained from descriptive statistics and visualisations, formulate hypotheses about your data. For example, you might hypothesise that high customer service interaction frequency is related to increased churn rates. These hypotheses can guide your subsequent analysis and model-building efforts.

Employing these methods can help you uncover critical insights from your data, guiding your understanding and informing your decision-making process as you prepare for more advanced data analysis stages.

# Best practices for cleaning and preparing data for machine learning models

Properly cleaning and preparing data is essential for building reliable and accurate machine learning models. This process ensures that the data used in model training is of high quality, which directly impacts the model's performance.

#### **Handling missing data**

Missing data is a typical problem that can significantly affect your model's accuracy. Here are a few strategies to address this problem:

* **Imputation:** Replace missing values with a statistical measure such as the mean, median, or mode of the column. For categorical variables, the most frequent category can be used.
* **Deletion:** Remove rows or columns with missing values, particularly if the proportion of missing data is small. However, this should be done cautiously to avoid losing valuable information.
* **Flagging:** Create a new binary column that flags whether data was missing in the original data set. This can help your model learn if the absence of data is itself informative.

#### **Outlier detection and treatment**

Outliers can skew the results of your machine-learning model. Use visualisation techniques like box plots or statistical methods to detect outliers. Once identified, you can:

* **Remove outliers:** If they are caused by data entry errors or are not relevant to the analysis.
* **Cap outliers:** Set a threshold beyond which data is capped. This technique minimises the influence of extreme values without removing data points entirely.

#### **Normalisation and standardisation**

Data normalisation and standardisation are techniques used to ensure that numerical features contribute equally to the model's learning process. These processes involve:

* **Normalisation:** Rescaling the values of numeric features to a common scale, typically [0, 1]. This is useful when features have different units or scales.
* **Standardisation:** Transforming data to have a mean of zero and a standard deviation of one. This process is particularly useful when the data follows a Gaussian distribution.

#### **Encoding categorical variables**

Machine learning models require numerical input, making it necessary to convert categorical data into numerical form. Common methods include:

* **One-hot encoding:** Creating binary columns for each category in a categorical feature. This method prevents the model from assuming any ordinal relationship between categories.
* **Label encoding:** Converting each category to a numerical value. This method is simpler but should be used with caution as it can imply an ordinal relationship where none exists.

#### **Feature engineering and selection**

Creating new features from the existing data (feature engineering) and selecting the most relevant features (feature selection) can significantly improve model performance. Techniques include:

* **Creating interaction features:** Combining two or more features to capture interactions.
* **Feature scaling:** Adjusting the range of features to ensure they contribute equally to the model.
* **Dimensionality reduction:** Using methods such as principal component analysis (PCA) to reduce the number of features, which can improve model performance and reduce overfitting.

By adhering to these best practices, you'll ensure that your data set is clean, well-prepared, and optimised for building effective machine learning models. This foundational work is crucial for achieving accurate and reliable predictions in your project.

Quiz

Question 1 of 3

* 1. Which of the following is the most critical step when identifying relevant data sources for predicting customer churn?

Correct! Identifying specific variables that influence customer behaviour and churn is crucial because it ensures that the data collected is directly relevant to the predictive model being developed, thereby improving its accuracy and effectiveness.

2. When performing EDA, why is it important to visualise data distributions and relationships between variables?

Correct! Visualising data distributions and relationships helps in identifying potential outliers and understanding the spread and central tendency of data, which are critical for making informed decisions during the data preprocessing and feature selection stages.

1. Which practice is most effective for ensuring that a machine learning model does not give undue weight to variables with larger numerical ranges?

Correct! Normalising or standardising numerical features ensures that variables with larger numerical ranges do not disproportionately influence the model's predictions, leading to more balanced and accurate outcomes.

# Data gathering and exploratory analysis

Now that you've been introduced to the scenario and the importance of this project, it's time to roll up your sleeves and get started. This task will challenge you to apply your data science skills in a real-world context, helping you connect theoretical knowledge and practical application.

#### **Data collection: start with relevance**

Your first step is to identify and gather relevant data that will provide insights into customer churn. Focus on data that can help you understand customer behaviour, such as demographics, transaction history, and customer service interactions. Remember, the goal is to collect data that is pertinent to the problem at hand. Approach this step with a critical eye, considering how each piece of data might contribute to understanding why customers are leaving.

#### **EDA: discover patterns and insights**

Once you've collected the data, the next step is to perform EDA. This phase is crucial as it helps you uncover patterns, identify anomalies, and understand the data's structure. Use visual tools such as histograms, scatter plots, and heat maps to explore relationships between variables. Pay close attention to trends that might indicate early signs of churn, such as decreased usage frequency or increased interaction with customer service.

EDA is not just a technical exercise; it's an opportunity to hypothesise the underlying causes of churn. For example, if you notice that customers who engage less with your digital services are more likely to churn, this insight could inform targeted retention strategies. Your analysis should be thorough and well-documented, providing a clear narrative that connects your findings to potential business actions.

#### **Data cleaning and preparation: ensuring quality**

The quality of your data affects the reliability of your predictive model. In this stage, focus on cleaning and preparing your data set. Handle missing values appropriately, either through imputation or removal, and ensure that all variables are in a consistent format. Normalising or standardising your data may be necessary, especially if the data includes variables with different scales. This step is about precision and care; small errors can lead to significant inaccuracies in your model.

As you work through these activities, keep the broader project goals in mind. Your findings from EDA and your cleaned data set will form the foundation for building a robust predictive model. This model will help SmartBank not only understand current churn rates but also anticipate and mitigate future risks, directly influencing customer retention strategies.

#### **Why accuracy and thorough understanding matter**

Your work in this task is not just about completing an assignment; it's about developing a deep, practical understanding of data science in action. Accuracy and thorough understanding are crucial, as the insights you derive will inform strategic decisions at SmartBank. The quality of your analysis could mean the difference between retaining valuable customers and losing them to competitors.

Approach each step with diligence and an analytical mindset. This is your opportunity to make a tangible impact on a real-world problem, honing your skills in the process. Embrace the challenge, knowing that your contributions are vital to the project's success.

# Task instructions

**Introduction**

In this task, you will take the first critical steps toward building a predictive model for customer churn. Your work will involve gathering relevant data, conducting EDA, and preparing the data set for model development. These activities are foundational for ensuring the accuracy and reliability of your subsequent analysis and predictions.

**Instructions**

**Identify and gather data:**

* Review the provided data sources and select those most relevant for predicting customer churn. Focus on key areas such as customer demographics, transaction history, and customer service interactions.
* Document your selection criteria and rationale for choosing each data set, ensuring that the data will provide meaningful insights into customer behaviour.

**Perform EDA:**

* Use statistical techniques and data visualisation tools to explore the data sets. Create visualisations such as histograms, scatter plots, and box plots to understand distributions, trends, and relationships between variables.
* Identify key features that may influence customer churn, paying special attention to patterns or anomalies that could be significant.

**Clean and preprocess the data:**

* Handle missing values by choosing appropriate methods such as imputation, removal, or flagging. Justify your chosen method based on the data and context.
* Detect and address outliers that could skew the analysis or predictions. Decide whether to cap, transform, or remove outliers based on their nature and potential impact.
* Standardise or normalise numerical features to ensure consistent scales across variables. This step is crucial for preparing the data for machine learning algorithms.
* Encode categorical variables using techniques like one-hot encoding to transform them into a numerical form appropriate for analysis.

**Deliverable:**

* **File submission:** Submit a comprehensive report detailing your data gathering, EDA, and data cleaning processes. The report should include:
  + A summary of the data sets selected and the rationale for their inclusion
  + Visualisations and statistical summaries from the EDA
  + A description of the data cleaning and preprocessing steps taken
  + The cleaned and preprocessed data set ready for model building

Ensure that your report is clear, concise, and well-organised, as it will be a key component of the project's success, guiding future analysis and model development.

# Nice work!

You just completed Task 1 of the Data Science Job Simulation

------------------------------------------------

Task 2: Building a machine learning model

-----------------------------------------

Task overview

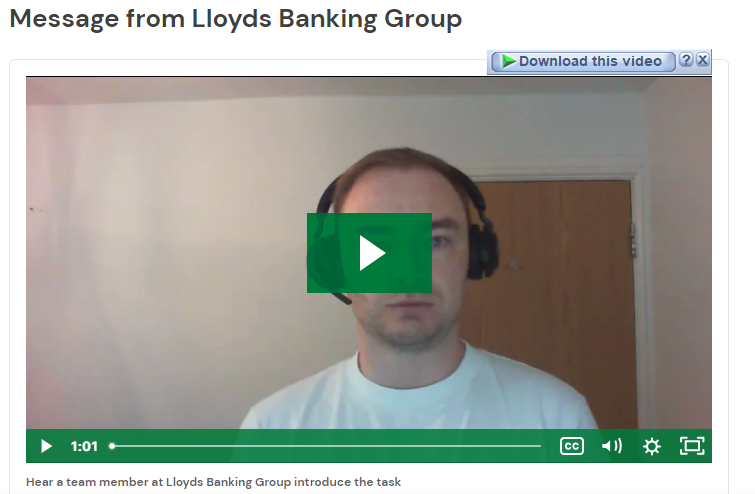
Excellent work on exploring and preparing the data in Task 1! Now, you'll leverage your analytical skills to build a predictive machine learning model, identifying key factors that influence customer churn and evaluating its performance for strategic decision-making.

**What you'll learn**

* Approaches to selecting appropriate machine learning algorithms
* Approaches to selecting and building machine learning models suited for classification tasks
* Techniques for suggesting ways to evaluate and measure the model’s performance

**What you'll do**

* Select an appropriate machine learning algorithm
* Build a model to predict customer churn
* Suggest ways to evaluate and measure the model’s performance



# Let's get started

You've already laid a solid foundation by exploring and preparing the data. Now, it's time to apply machine learning techniques to build a predictive model for customer churn, a crucial step in enhancing SmartBank's customer retention strategies.

In this task, your focus will shift from data preparation to model development. The challenge lies in selecting the right machine learning algorithm and fine-tuning it to accurately predict which customers are at risk of leaving. This model will provide actionable insights, enabling the team to develop targeted interventions to retain valuable customers.

Li, your mentor, has emphasised the importance of accuracy and precision in this phase. "The insights from this model will drive our retention strategies. It’s crucial that we build a model that not only predicts churn but also provides clear indicators of why customers are leaving," she explains. This guidance underscores the practical implications of your work; the model you build must be both accurate and interpretable.

Your task involves choosing an appropriate algorithm, training the model, and evaluating its performance. Remember, the goal is to create a model that can be easily understood and acted on by business stakeholders. As you work through this task, consider the factors that could influence churn, such as spending habits, service usage, and demographic characteristics.

This is a chance to showcase your data science expertise in a real-world scenario. Your efforts will not only enhance your skills but also contribute significantly to the team's understanding of customer behaviour. As you begin, keep in mind the practical impact of your model on SmartBank's strategic decisions. Let's get started and bring your analysis to life!

# Approaches to selecting appropriate machine learning algorithms

Selecting the right machine learning algorithm is crucial for building a robust predictive model. Given the complexity of customer churn prediction, where the target variable is categorical, you need to consider several factors that influence the choice of the model. Here are some approaches to help guide your selection.

#### **Understanding the problem type and data characteristics**

In churn prediction, you're dealing with a **binary classification problem**. Key considerations include:

* **Imbalance in the data set:** Customer churn data sets often have an imbalance, where the number of churned customers is significantly less than non-churned. Techniques like **resampling**, **SMOTE (synthetic minority over-sampling technique)**, or **adjusted class weights** in algorithms are crucial for handling this imbalance effectively.
* **Feature engineering:** Advanced feature engineering techniques, such as **interaction terms**, **polynomial features**, and **dimensionality reduction (e.g., principal component analysis)**, can significantly influence the performance of algorithms, especially those sensitive to multicollinearity and high-dimensional spaces, like logistic regression and support vector machines (SVMs).

#### **Algorithm selection and considerations**

* **Logistic regression:** Preferred for its simplicity and interpretability, logistic regression can be enhanced with **regularisation techniques (L1, L2)** to prevent overfitting, especially in high-dimensional data sets.
* **Decision trees and random forests:** These are powerful for capturing non-linear relationships and interactions between features. Random forests, an ensemble of decision trees, provide robustness against overfitting and allow for **feature importance analysis**, which can be crucial in understanding which factors contribute most to churn.
* **SVMs:** Effective in high-dimensional spaces and when the decision boundary is not linear. The use of **kernel tricks (e.g., RBF, polynomial)** allows SVMs to handle non-linear relationships, but they require careful tuning of hyperparameters such as **C (regularisation parameter)** and **gamma**.
* **Neural networks:** While potentially offering high accuracy, especially with complex data patterns, they require large amounts of data and computational power. Techniques like **dropout**, **batch normalisation**, and **early stopping** are essential to prevent overfitting.

#### **Model evaluation and tuning**

* **Cross-validation:** Advanced cross-validation techniques, such as **stratified k-fold**, ensure that each fold has a representative distribution of the target class, crucial for imbalanced data sets.
* **Hyperparameter tuning:** Employ **grid search** or **random search** for systematic exploration of the hyperparameter space. For more efficient optimisation, consider using **Bayesian optimisation** or **automated machine learning (AutoML)** tools.

#### **Scalability and practical considerations**

* **Model deployment:** Consider the model's scalability and integration into the business workflow. This includes **real-time prediction capabilities**, ease of updating the model with new data, and computational efficiency.
* **Interpretability vs. accuracy trade-offs:** In practice, balancing interpretability with predictive power is often necessary, especially when model decisions need to be transparent to stakeholders.

By delving into these advanced considerations, you'll be better equipped to select and fine-tune machine learning algorithms that are both accurate and aligned with the practical needs of the business context in which they will be deployed.

(page 4)

# Approaches to selecting and building machine learning models for classification tasks

Building a machine learning model for classification tasks, such as predicting customer churn, requires a deep understanding of both the algorithmic foundation and the practical nuances of implementation. Here are some advanced approaches to guide you through this process.

#### **Feature selection and engineering**

* **Dimensionality reduction:** Techniques such as **principal component analysis (PCA)** or **t-distributed stochastic neighbour embedding (t-SNE)** can be used to reduce the feature space, mitigating the curse of dimensionality and enhancing model performance.
* **Feature importance analysis:** Algorithms like random forests provide intrinsic measures of feature importance, which can guide the selection of the most predictive features. This step is crucial for simplifying the model and improving interpretability without sacrificing accuracy.
* **Interaction terms and polynomial features:** Introducing interaction terms and polynomial features can capture non-linear relationships between variables, which are often missed in linear models. This is particularly useful in models like logistic regression, where extending the feature space can significantly enhance predictive capability.

#### **Model selection and evaluation**

Choosing the right model involves balancing several factors:

* **Algorithm suitability:** While logistic regression and decision trees offer simplicity and interpretability, they may lack the predictive power of more complex models like **gradient boosting machines (GBMs)**, **XGBoost**, or **neural networks**. The choice often depends on the trade-off between model performance and explainability.
* **Model evaluation metrics:** In the context of imbalanced data sets, traditional metrics like accuracy are often misleading. Use metrics such as **precision, recall, F1-score,** and **ROC-AUC** to get a more accurate picture of model performance. Additionally, the **confusion matrix** provides detailed insights into the true positives, false positives, true negatives, and false negatives, which are critical for understanding model behaviour.

#### **Advanced model tuning techniques**

Optimising model performance involves fine-tuning hyperparameters:

* **Grid search and random search:** These methods are standard for hyperparameter optimisation but can be computationally expensive. Grid search is exhaustive, covering all combinations of specified hyperparameters, while random search samples a wide range but in a more computationally efficient manner.
* **Bayesian optimisation:** For more efficient hyperparameter tuning, Bayesian optimisation offers a probabilistic approach to finding the optimal parameters, often outperforming traditional methods in terms of both accuracy and computational cost.
* **Cross-validation:** Use **stratified k-fold cross-validation** to ensure that each fold has the same proportion of classes as the original data set, which is crucial for imbalanced classification tasks. This approach helps in validating that the model generalises well to unseen data.

#### **Model implementation and scalability**

Once a model is selected and tuned, consider its deployment and scalability:

* **Pipeline integration:** Incorporate the model into a robust data pipeline, ensuring it can handle real-time data streams and integrate seamlessly with existing systems. This includes automating data preprocessing, model prediction, and output generation.
* **Model monitoring and maintenance:** Post-deployment, continuously monitor model performance to detect drifts in data distribution or declines in accuracy. Implementing **version control** for models, along with retraining strategies, ensures the model remains accurate and relevant as new data becomes available.

By integrating these advanced techniques, you'll build a classification model that is not only accurate and robust but also scalable and maintainable, ensuring long-term value for the business.

# Techniques for suggesting ways to evaluate and measure the model’s performance

Evaluating and measuring the performance of a machine learning model, especially in classification tasks like predicting customer churn, is crucial for understanding its effectiveness and reliability. Here are some advanced techniques and metrics to ensure comprehensive evaluation.

#### **Choosing the right evaluation metrics**

Selecting appropriate metrics depends on the specific characteristics of the data set and the business objectives:

* **Precision and recall:** These metrics are particularly important in imbalanced data sets where false positives and false negatives carry different costs. **Precision** measures the proportion of true positive predictions among all positive predictions, while **recall** measures the proportion of true positives identified out of all actual positives.
* **F1 score:** The F1 score balances precision and recall, offering a single metric that accounts for both false positives and false negatives. This is particularly useful when the costs of these errors are similar.
* **ROC-AUC (receiver operating characteristic - area under curve):** The ROC-AUC score evaluates the trade-off between true positive rates and false positive rates across different threshold settings. A higher AUC indicates better model performance across various decision thresholds.
* **Confusion matrix:** This matrix offers a detailed breakdown of the true positives, false positives, true negatives, and false negatives. It is a fundamental tool for understanding model performance, especially in terms of misclassification types.

#### **Model calibration and validation**

* **Calibration curves:** To assess how well predicted probabilities align with actual outcomes, use calibration curves. These curves compare predicted probabilities with actual outcome frequencies, helping to adjust the model to improve probability estimation.
* **Cross-validation:** Beyond simple training-validation splits, **k-fold cross-validation** ensures that the model's performance is consistently evaluated across different subsets of the data. This technique reduces the likelihood of overfitting and ensures that the model generalises well to unseen data.
* **Bootstrapping:** This statistical method involves repeatedly resampling the data set with replacement to estimate the distribution of model performance metrics. Bootstrapping provides insights into the variability and robustness of the model's predictions.

#### **Post-model analysis**

* **Feature importance and SHAP values:** Understanding why a model makes certain predictions is crucial, especially in business contexts where decisions must be justified. **Feature importance** metrics in models like random forests and **SHAP (SHapley Additive exPlanations)** **values** provide insights into how each feature contributes to the model’s decisions.
* **Error analysis:** Conduct a thorough analysis of the model's errors, focusing on cases where the model performs poorly. This analysis can reveal data patterns that the model misses, leading to insights for further feature engineering or model adjustments.
* **Business impact analysis:** Beyond statistical metrics, evaluate the model's performance in terms of business outcomes. For example, measure the impact of the model on customer retention rates or revenue. This analysis helps assess the model's practical value.

#### **Continuous monitoring and reassessment**

* **Model drift detection:** Implement systems to detect model drift, which occurs when the data distribution changes over time, leading to a decline in model performance. Techniques like monitoring prediction probabilities or feature distributions can help the early detection of drift.
* **Retraining strategies:** Based on performance monitoring, establish criteria for retraining the model. This could involve periodic retraining or retraining triggered by specific performance thresholds or detected drifts.

By employing these advanced techniques for model evaluation and measurement, you ensure that the predictive model not only performs well statistically but also aligns with business goals, providing actionable insights and reliable predictions.

Quiz

1. Which factor is most critical when choosing a machine learning algorithm for predicting customer churn in a data set with a high class imbalance?

The availability of techniques to handle class imbalance, such as adjusting class weights or using SMOTE

Correct! In cases of class imbalance, the most critical factor is the ability of the algorithm to effectively handle this imbalance, which can be achieved through techniques like adjusting class weights or using oversampling methods such as SMOTE.

1. Why is it important to use techniques like principal component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE) in feature selection?

To reduce dimensionality and capture underlying patterns that may not be apparent in the original feature space

Correct! PCA and t-SNE are used to reduce dimensionality, which helps in capturing underlying patterns and structures in the data that may not be apparent in the original feature space, thereby improving model performance and interpretability.

1. In the context of an imbalanced data set for churn prediction, why might the ROC-AUC score be preferred over accuracy as a performance metric?

Accuracy can be misleading in imbalanced data sets, as it may not reflect the true performance on minority classes

Correct! In imbalanced data sets, accuracy can be misleading as it may not reflect the model's performance on minority classes. ROC-AUC provides a better assessment by considering the trade-offs between true positive and false positive rates, giving a more comprehensive view of model performance across various threshold settings.

# Crafting the predictive blueprint

As you delve deeper into the intricacies of data science, Task 2 offers a critical opportunity to influence the strategic direction of Lloyds Banking Group. Your work in this phase will pivot from EDA to the construction of a predictive model that could significantly shape the bank's customer retention strategies.

The urgency and importance of this task cannot be overstated. Lloyds has seen a subtle but concerning trend in customer attrition, which, if not addressed, could have substantial implications for its market standing and profitability. This project is a real-world application where your findings could directly influence business decisions and outcomes.

**Team collaboration and strategic impact**

Working closely with Li and the Data Science & Analytics team, you are stepping into a role where your technical expertise merges with strategic business needs. The team has identified key areas where predictive insights could allow for proactive interventions, thus reducing churn and enhancing customer loyalty. Li underscores the strategic importance of your task, noting, "Our ability to predict churn allows us to personalise our engagement strategies, tailoring our approach to meet the needs and preferences of our customers. This is critical for maintaining a competitive edge."

**Integrating analytical insights into business strategies**

Your task involves selecting the most appropriate machine learning algorithm, which balances predictive accuracy with interpretability. This balance is crucial; while complex models may offer high accuracy, they can be challenging to explain to non-technical stakeholders who need to trust and act on these insights. Therefore, the choice of algorithm must consider not only the statistical performance but also the business context and usability.

Building the model is where your analytical skills will shine. Using the preprocessed data from Task 1, you'll train the model to identify patterns indicative of potential churn. This involves iterating on various models, tuning hyperparameters, and validating the model to ensure it generalises well to new data.

**Evaluating and communicating model performance**

Beyond building the model, it's essential to suggest robust evaluation metrics. The goal is to provide a comprehensive view of the model's performance, highlighting its strengths and identifying any limitations. Metrics like precision, recall, and F1 score will be key, especially in the context of imbalanced data sets where simple accuracy might be misleading. Furthermore, explaining the model's predictions through feature importance or other interpretability tools will be vital for aligning the model's outputs with business decisions.

**Delivering actionable insights**

The culmination of your work will be a detailed report. This report should present the technical aspects of the model and translate these findings into actionable business insights. Your ability to communicate complex data science concepts in a clear and actionable manner will be crucial in ensuring that the strategic implications of your work are understood and implemented by the business.

This task is a critical component of Lloyds' broader strategy to harness data-driven insights for business growth. Your contributions will play a pivotal role in shaping how the bank understands and responds to customer needs, ultimately driving customer satisfaction and loyalty. As you embark on this task, remember that your work has the potential to make a significant impact, both analytically and strategically.

# Task instructions

**Introduction**

In this task, you will focus on developing a robust machine learning model to predict customer churn. Your objective is to select an appropriate algorithm, train and validate the model, and propose evaluation metrics that will help assess its performance. This task is pivotal for providing actionable insights that can inform business strategies at Lloyds Banking Group.

**Instructions**

**Select an appropriate machine learning algorithm:**

* Review the characteristics of the data set and the nature of the churn prediction problem.
* Consider algorithms such as logistic regression, decision trees, random forests, gradient boosting machines, or neural networks.
* Choose an algorithm that balances accuracy and interpretability, suitable for the business context.

**Build and train the model:**

* Use the preprocessed data set from Task 1 to train your chosen model.
* Implement techniques like cross-validation to ensure the model generalises well to unseen data.
* Perform hyperparameter tuning to optimise the model’s performance.

**Evaluate model performance:**

* Select appropriate metrics to evaluate the model's performance, such as precision, recall, F1 score, ROC-AUC, and confusion matrix analysis.
* Consider the implications of each metric in the context of imbalanced data sets, ensuring that the evaluation provides a comprehensive view of the model's effectiveness.

**Suggest ways to improve and utilise the model:**

* Provide recommendations on how the model can be used by the business to identify at-risk customers and develop retention strategies.
* Discuss any potential improvements or adjustments to the model that could enhance its accuracy or applicability in different business scenarios.

**Deliverable:**

* **Report submission:** Compile a comprehensive report that includes:
  + A detailed description of the selected algorithm and the rationale behind its choice.
  + The trained model, along with performance metrics and evaluation results.
  + Suggested ways to utilise the model's predictions for business decision-making and potential areas for improvement.

Ensure that your report is clear, concise, and well-organised, effectively communicating both the technical aspects of the model and its practical applications for the business. This report will be a critical tool for stakeholders to understand and leverage the predictive insights generated by your model.