**Template – Identify Use Case**

**Milestone 1: What is the goal of the use case?**

|  |
| --- |
| *Describe what the use case is and the goal that you aim to achieve* |

**Building blocks**

|  |
| --- |
| How will you measure your goal?   * Percentage increase revenue * Improved customer experience ratings * Increased efficiency (reduction of time and/or cost, enhanced quality) * Realize competitive advantage (increased market share, improved market position) * Reduce business risks |

|  |
| --- |
| What is the reason to consider AI in this case?   * What is the current situation and the key problem? * What is the desired situation and why would this be a solution? |

|  |
| --- |
| Who will this project impact and in which way?   * Who are the most important stakeholders? (Internal and external to KPN.) * What are the needs of the intended audience? |

|  |
| --- |
| What are the benefits and the risks of approaching this case with AI?   * What are benefits? * What are risks? |

**Milestone 2: Define the question you want to answer**

|  |
| --- |
| *Formulate the question that the AI application must answer:* |

**Building blocks**

|  |
| --- |
| What is the business question you would like to answer? We want to…   * Describe... * Discover... * Predict... * Prescribe...   Is there only one question or does the use case break down into multiple questions?   * Consider the order of questions and nesting (fractal approach) |

|  |
| --- |
| What are the boundaries of the question?   * Where does the model apply / not apply?   + Example: Europe yes, Asia no. Medical industry: yes, manufacturing industry: no. * Features within the dataset that represent limits?   + Example: Gender, age, country * Product or service: For which product or service is the model? |

|  |
| --- |
| What are the constraints of the question?   * Timeliness: how fast or frequent can you train and evaluate the model? * Legal: what ethical or legal dilemmas can you expect? * Interpretability: Is it important that people understand the trade-offs (White box) or is explanation less important and more result-oriented? (Black box) * Technical: data complexity, data size, data storage or speed. |

**Milestone 3: What is the problem type?**

|  |
| --- |
| *Consider and identify the nature of the problem.* |

**Building blocks**

|  |
| --- |
| **Supervised Learning**  An approach to let an algorithm learn how data inputs are related to (predetermined) outputs. The algorithm is trained based on examples in which the desired output is known, hence the term ‘supervised’. There are many different algorithms within this category, but on a general level they are used to either [1] categorize data into a given number of classes (‘classification’) or [2] predict a (numerical) value, usually referred to with the term ‘regression’. In both cases there is labeled output to train the algorithm with, but the difference is the type of output the algorithm needs to model (classes or values).   1. Classification: Categorize data into a given number of classes (for prediction or interpretation). 2. Regression:    1. Predicts a target value given related data features.    2. Estimates the effect of a factor(s) on the target. |

|  |
| --- |
| **Unsupervised learning**  Bottom-up approach for finding patterns in the data. In contrast to supervised learning (classification and regression above), the desired output is not specified by the user but rather the throughput is specified c.q. how the computer may process the data to find certain patterns and outputs. Types of unsupervised learning that are used relatively often:   1. Clustering: used to find distinct groupings of data. This can be applied on any type of data (numbers, text, pictures, …) 2. Dimensionality reduction: used to reduce the number of dimensions. This can be applied for example when wanting to visualize a multidimensional vector in 2D or 3D, or to reduce many questions in a survey to a few underlying concepts. 3. Process mining: used to describe process data based on the actual data. 4. Collaborative filtering: used to filter data on an individual level based on collective patterns. |

|  |
| --- |
| **Reinforcement Learning**  Used to evaluate performance based upon an intended outcome and reward/penalize accordingly. This problem-type is important for algorithms that will carry out tasks that mimic human behavior. |

**Milestone 4: Choose a solution method for the problem type**

**4.1** **Classification**

|  |
| --- |
|  |

**Building blocks**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Explanation** | **Example Use Case** |
| **Logistic Regression** | Estimates the relationship between a binary target (e.g. true or false) and one or more features by providing the probability of either option between 0 and 1 | Credit scoring: Risky or not  Hotel management: Guest will cancel or not |
| **k-Nearest Neighbours** | Uses the proximity of data points with each other to create classes with the lowest distance with points in the same class and maximum distance from another class | Medicine: Has cancer, requires more attention, and no cancer (More than 2 classes)  Product offerings: group similar products together in product groups |
| **Decision Tree or Forest** | Repeatedly subsets observations based on multiple features provided and probability of that difference being accurate | Market research: Categorize consumers based on a target/goal  Management: Helps to guide decision making in a wide variety of contexts |
| **Neural Network** | A series of algorithms that attempt to un-cover a under-lying relationship in data though a structure that mimics human neurons | Motion: A computer detecting classes of actions  Vision: Augmented reality detecting specific objects  Robotics: Roomba choosing the most efficient direction (left, right, back, forward) |
| **Support Vector Machine** | A support vector machine tries to locate the optimal hyperplane (boundary between data groups) that is used to separate and classify labelled data | Detection systems: Email spam detection, Fraud detection  Face detection: Can be used to distinguish between faces and not, placing squares around faces. |
| **Naïve Bayes** | Assumes features are independent from each other, using each as an individual classify for a classification task | Text analysis: Can use text inputs to predict topics or underlying concepts  Real-time predictions: Since this algorithm is efficient, it can be used in-real time and effectively. |

**4.2** **Regression**

|  |
| --- |
|  |

**Building blocks**

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Explanation** | **Example Use Case** |
| **Linear Regression** | Estimates the relationship between one target and multiple features using a straight line | Financial Performance: Predicting profitability or promotional effectiveness  Market Research: Interpreting Consumer behaviour through data recorded about their actions |
| **Time Series Forecasting** | A forecasting method that attempts to predict future outcomes based on historical outcomes, and the consistent transfer of dynamics from relevant predictors | Financial Performance: Historical performance used to forecast future performance, allowing for trend investigation |
| **Stochastic Gradient Descent (Optimization)** | Algorithm iteratively approximates the gradient at a single step in determining classifiers until it reaches the lowest possible error | Cost reduction: minimize costs given data features.  Revenue improvement: Profit maximization using factors that provide information on products, demand, supply, etc. |
| **Neural Network** | A series of algorithms that attempt to un-cover a under-lying relationship in data though a structure that mimics human neurons | Management: Predict sales, profit, revenue, cost figures  Vision: Augmented reality detecting measurement of objects  Robotics: Roomba choosing the most efficient direction (360 degrees) |
| **Decision Tree or Forest** | Repeatedly subsets observations based on multiple features provided and probability of that difference being accurate | Market research: Categorize consumers based on a ranges of ages, incomes or other numeric features |

**4.3** **Unsupervised learning**

|  |
| --- |
|  |

**Building blocks**

|  |  |  |
| --- | --- | --- |
| **k-Means Clustering** | This algorithm tries to find clusters based on a pre-defined number of k clusters using the data points with the closest means to each other | Market research: customer segmentation (Customer metrics)  Supply-chain management: Group suppliers based on relationship status, or risk rating |
| **F-P Growth (Pattern or Association Algorithm)** | This algorithm develops a frequency tree or pattern which highlights the most frequent association patterns among features in a dataset | Market research: Discover brand or product associations, based on how the algorithm connects them |
| **Process discovery** | This algorithm translates log data into a process model | Market research: discover how customers interact with the organisation, what steps are taken in the customer care, what ‘journeys’ people take. |
| **k-Modes Clustering** | This algorithm is a variation of k-means which deals specifically with non-numeric data, by instead clustering observations based of closest means, observations are grouped based on the mode (how frequently an observation occurs) of those categories | Market research: Allows for clustering based on categories, which can be used to protect privacy and data by not showing specific measurements |
| **Gaussian Mixture Model** | GMM assumes that there is a mixture of k components, each with probability distributions and k mixture weights, which are then used to group observations | Market research: Customer segmentation in the event that sub-populations cannot or have not been identified |
| **Principal Component Analysis (PCA)** | This is a data transformation technique that alters the data to a new coordinate system such that correlation between similar dimensions can be grouped or removed | Market research: check if correct concepts were measured  Customer journey: Streamline information gathering by reducing questions asked or surveys used  General: visualize data in 2D or 3D where the data you want to visualize is multidimensional (e.g. 40-dimensional) |
| **Latent Dirichlet Analysis (LDA)** | A technique for natural language processing that counts occurrences of words that relate to pre-defined topics within a set of documents | Natural language processing: Discover important topics and their frequency in blocks of text  Natural language processing: Aggregate texts blocks used by a group of people to estimate if their opinion is positive or negative in that context |
| **Isolation Forest (Anomaly Detection)** | The algorithm uses decision forests as a base, but then splits data to ensure only 1 data point is filtered out at each split. | Anomaly detection: Can be used to detect unique issues or obstacles recorded within data, like specific cases of mis-inputs or events |
| **Collaborative filtering** | This algorithm uses data on a group level to filter data for an individual entity. | This algorithm is widely used in recommender systems where an individual gets a recommendation based on the ‘consumption’ of relatively similar customers. |

**Milestone 5: Data Understanding**

|  |
| --- |
| Describe the data required for the AI solution |

**Building blocks**

|  |
| --- |
| What data is required to address the business question?   * What data is available, do you have access to this? * What data is not available, and how will we acquire it? |

|  |
| --- |
| What are the features required within the data?   * What is required to identify these features? * Structured / Semi- / Unstructured |

|  |
| --- |
| How should this data be interpreted?   * Syntax and semantic? Labelling? * Does the data reflect the essence of the environment it represents? |

|  |
| --- |
| Who owns the data?   * Access and the use of the data, are there constraints around the data? Example: GDPR |

|  |
| --- |
| How will biases be identified within the data? |

**Milestone 6: Data Preparation & Validation**

*How would you prepare your data?*

* Clean data (visualize data distributions)
* Generate new attributes required for analysis (also called feature engineering)
* Integrate data (Merge and aggregate)

*How would you validate your data?*

* Is the size of data enough compared to the population?
* Is there a sample bias?
* Check whether the data has a correlation with the objective? (Correlation does not mean causality but could be an indicator)
* Check whether the data contains a reliable feature to predict and verify?
* Does the data need transformation to process? What coding techniques fit?
* Others...

**Milestone 7: Statistical Evaluation**

**Key Statistical metrics**

|  |
| --- |
| *Choose minimal two metrics (if applicable, otherwise describe subjective factors to evaluate un-supervised outcomes).* |

**Building blocks**

|  |  |  |
| --- | --- | --- |
| **Classification Metric** | **Calculated** | **Explanation** |
| Confusion matrix | Actual vs Predicted | Diagram showing: True Positives, True Negatives, False Negatives, False Positives. |
| Accuracy | (True Positives + True Negatives) / Total events | Number of correct predictions compared to total predictions. |
| Precision | True Positives / (True Positives + False Positives) | Number of correct positive predictions (ratio true versus predicted positives) |
| Recall | True Positives / (True Positives + False Negatives) | Number of correct predictions for the positives (ration true versus actual positives) |

|  |  |  |
| --- | --- | --- |
| **Regression Metric** | **Calculated** | **Explanation** |
| Mean Absolute Error (MAE) | For each prediction P – A = Diff | Sum of absolute difference between Actual and Predicted. |
| Mean Square Error (MSE) | For each prediction P – A = Diff2 | Sum of absolute difference squared between Actual and Predicted. |
| R Squared (R2) | MSE / Constant Baseline = R2 | How much better than constant baseline? 0 = not better , 1 = perfect |

|  |
| --- |
| Clustering and Dimensional Reduction do not contain metrics, because these problem types don’t contain a target. These problem types are used to generate new insights from data. |

**Milestone 8: Business Evaluation**

**How would you evaluate if the solution is *Fit for purpose*?**

|  |
| --- |
| What are the key criteria to judge the solution? Does the solution answer the question? |

**How would you evaluate if it is ethical and legal?**

|  |
| --- |
| *Consider the EU Trustworthiness framework and upcoming legal requirement* |

|  |
| --- |
| Would the use of an AI solution influence the environment in an adverse manner?  E.g. consider the effect of a False Negative versus a False positive in the environment |

