# MSc Artificial Intelligence and Data Science

# Capstone Project Proposal

|  |  |  |
| --- | --- | --- |
| **Project Title** | | |
| **RiskTracker: A Text classification framework for news article tagging and risk type classification.** | | |
| **Apprentice Title** | **Apprentice First Name** | **Apprentice Last Name** |
| Mr | Dylan | Viswambaran |
| **Employer Name** | **Mentor/Line Manager Name** | **Mentor/Line Manager Email** |
| Bank of England | Paul Boyle | [Paul.Boyle@bankofengland.co.uk](mailto:Paul.Boyle@bankofengland.co.uk) |
| **Academic Supervisor Name** |  |  |
| TBC |  |  |

|  |  |
| --- | --- |
|  | You can make suitable adjustments to this template as required. The information in the blue boxes is here to help you understand what should be included in each section. |

## Project Background and Rationale

|  |  |
| --- | --- |
|  | Describe how this project came about, who is involved, the purpose of the project, and the expected benefits (value to the organisation). |

|  |
| --- |
| The rapid growth of digital media and the abundance of news articles makes it increasingly challenging to extract meaningful insights from the vast amount of news available. In light of this I’ve identified a need to develop an assisted process for supervisors that can tag and classify news articles for topics of interest such as specific risks (credit, market, liquidity, fraud, fines and so on). This project would involve collaboration between data scientists and domain specialists such as supervisors to enhance accuracy of the final product. This produce should enhance decision making by serving as an early warning indicator and allow decisions to be made sooner than currently possible |

## Project Scope

|  |  |
| --- | --- |
|  | Project scope defines the boundaries of a project. Think of the scope as an imaginary box that will enclose all the project elements/activities. It not only defines what you are doing (what goes into the project), but it sets limits for what will not be done as part of the project (what doesn’t fit in the project). Scope answers questions including what will be done, what won’t be done, and what the result will look like.  Refer to the competencies you will demonstrate in this project (in section 5). |

|  |
| --- |
| Data collection – gathering a diverse and comprehensive dataset of news article.  Model development – developing a robust text classification model which will be trained to tag and classify news articles.  Feature engineering – exploring various techniques which can include bag of words, word embeddings, transformers or zero shot classification.  Testing and validation – testing and validation will be conducted to ensure the accuracy and reliability of the text classifier. Involves evaluating performance against a separate test dataset and comparing the outputs with human labelled annotations.  The project scope will not include:   * News article collection infrastructure: the project will not address using any specific news article APIs or sources but will focus on the development of the text classification model which can then be applied to the preferred news article supplier when chosed. * Real time news monitoring: this pipeline will be built to focus on building a robust model to use rather than real time news monitoring. * Multi lingual support: the project will focus on news article written in English * Continuous model improvement: whilst the models will involve iterations of model training and evaluation it will not include continuous improvement upon deployment. These will take place in future phases of the project. |

## Aim(s) and Objectives

|  |  |
| --- | --- |
|  | Describe the overall aim of your project and the supporting objectives that support the development. For example, your aim might be:   * To develop a system that facilitates the information sharing between the marketing team and operations teams   With the following objectives:   * To identify the requirements of the marketing team and operations teams * To identify the appropriate system development methods and techniques that can be used for the project * To develop a system based on the agreed requirements and development methods * To implement and evaluate the developed system |

|  |
| --- |
| Aims:   * Develop a text classification model that can accurately tag and classify news articles based on risk types and associated terms of interest. * Enhance and improve decision making for supervisors by providing them with timely and relevant risk related information extracted from news articles * Act as an early warning indicator to proactively identify areas of interest particularly around specific risks reported in news articles.   Objectives:   * Collect and process a diverse and comprehensive dataset of news articles suitable to train and test the model * Explore and implement appropriate feature engineering techniques to extract relevant features from news text. * Train and evaluate the models using the collected dataset to optimise its performance and accuracy * Integrate the model into a user friendly application with supporting documentation on the model specifics. |

## Deliverable(s) and Success Criteria

|  |  |
| --- | --- |
|  | Describe your project deliverables and how they can be measured and considered achieved and successful. For example, the deliverables can be:   * Systems requirements documentation * A new system that meets the user requirements   The success criteria could be:   * Ability to allow both internal and external users to access the application without downloading any software * Ability to interface with the existing data warehouse application * Ability to incorporate automated routing and notifications based on existing business rules |

|  |
| --- |
| Text classification Model System: The primary deliverable of this project is a trained text classification model capable of accurately tagging and classifying news articles into risk types.  User interface: An interface to be created that receives news article inputs and outputs relevant risk tags. It should be user friendly, reliable and meet performance requirements.  Documentation: Comprehensive documentation including technical specifications, instructions and troubleshooting guides.  Success Criteria:  Accuracy of classification model: The success of the project will be measured by the accuracy of the resulting model tagging and classifying articles into the correct categories and on a consistent basis.  System performance and reliability: The performance and reliability of the model. It should be able to handle a reasonable volume of news articles within acceptable response times and demonstrate stability and resilience in real world usage.  Achievement of Project Objectives: This includes the development of the text classification model , the enhanced risk monitoring, streamlined information processing, improved/faster decision making and an establishment of a form of an early warning indicator. |

## Competencies to be Demonstrated

|  |  |
| --- | --- |
|  | The tables list the core and specialist competencies to be demonstrated via the project. In the ‘how will it be met?’ cells, please describe how the apprentice will meet each requirement through the proposed project. |

|  |  |
| --- | --- |
| Core competencies (applicable for all specialisms) | How will it be met? |
| K1 How to use AI and machine learning methodologies such as data-mining, supervised/unsupervised machine learning, natural language processing, machine vision to meet business objectives |  |
| K3 How to apply advanced statistical and mathematical methods to commercial projects |  |
| K5 How to design and deploy effective techniques of data analysis and research to meet the needs of the business and customers |  |
| K6 How data products can be delivered to engage the customer, organise information or solve a business problem using a range of methodologies, including iterative and incremental development and project management approaches |  |
| K13 How to identify the compromises and trade-offs which must be made when translating theory into practice in the workplace |  |
| K14 The business value of a data product that can deliver the solution in line with business needs, quality standards and timescales |  |
| K23 The use of different performance and accuracy metrics for model validation in AI projects |  |
| K26 The scientific method and its application in research and business contexts, including experiment design and hypothesis testing |  |
| K28 How to communicate concepts and present in a manner appropriate to diverse audiences, adapting communication techniques accordingly |  |
| S2 Independently analyse test data, interpret results and evaluate the suitability of proposed solutions, considering current and future business requirements |  |
| S3 Critically evaluate arguments, assumptions, abstract concepts and data (that may be incomplete), to make recommendations and to enable a business solution or range of solutions to be achieved |  |
| S4 Communicate concepts and present in a manner appropriate to diverse audiences, adapting communication techniques accordingly |  |
| S5 Manage expectations and present user research insight, proposed solutions and/or test findings to clients and stakeholders. |  |
| S7 Work autonomously and interact effectively within wide, multidisciplinary teams |  |
| S9 Manipulate, analyse and visualise complex datasets |  |
| S10 Select datasets and methodologies most appropriate to the business problem |  |
| S11 Apply aspects of advanced maths and statistics relevant to AI and data science that deliver business outcomes |  |
| S15 Identify, develop, build and maintain the services and platforms that deliver AI and data science |  |
| S17 Consistently implement data curation and data quality controls |  |
| S18 Develop tools that visualise data systems and structures for monitoring and performance |  |
| S22 Apply scientific methods in a systematic process through experimental design, exploratory data analysis and hypothesis testing to facilitate business decision making |  |
| S24 Apply research methodology and project management techniques appropriate to the organisation and products |  |
| S25 Select and use programming languages and tools, and follow appropriate software development practices |  |
| S27 Analyse information, frame questions and conduct discussions with subject matter experts and assess existing data to scope new AI and data science requirements |  |
| B2 Reliable, objective and capable of independent and team working |  |
| B6 Is comfortable and confident interacting with people from technical and non-technical backgrounds. Presents data and conclusions in a truthful and appropriate manner |  |

## Conclusion

|  |  |
| --- | --- |
|  | Define the expected benefits achieved from using the project, the expected challenges faced during the life cycle of the project and the expected theoretical and empirical contribution. |

|  |
| --- |
|  |

## Software Application Requirements

|  |  |
| --- | --- |
|  | Identify the software applications and tools to be utilised in the project, e.g. R Studio, SQL, Java, JavaScript, Python, Excel, PowerPoint, Word, Power BI, Tableau, QlikView, Qlik Sense, etc. |

|  |
| --- |
| Python, RStudio |

## Project Plan/Schedule

|  |  |
| --- | --- |
|  | Describe what the high-level timeline/schedule will be to plan, design, develop and deploy the project. Generally, by when do you expect this project to be finished? When should the key milestones be achieved? You can consider using a Gantt Chart. |

|  |
| --- |
|  |

## Apprentice Agreement

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Position/Role in the organisation | Signature  (e-signature or scanned wet signature accepted. Typed signature NOT accepted) | Date |
| Dylan Viswambaran | Data Scientist |  | 07/07/2023 |
| *By signing this proposal you are agreeing that you have read and will adhere to the*[*University's Ethical code of conduct*](https://policies.northeastern.edu/policy101/)*.* | | | |

## Mentor/Line Manager Approval

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Position/Role in the organisation | Signature  (e-signature or scanned wet signature accepted. Typed signature NOT accepted) | Date |
|  |  |  |  |

## Appendix

|  |  |
| --- | --- |
|  | Include the relevant supporting information in this section. Remove this section if not required |

­­