**EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Mark Brackenrig (12964298) |
| **Project Name** | Assignment 1A |
| **Date** | 07/02/2021 |
| **Deliverables** | <notebook name>  <model name>  <other> |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | The goal of this project is to predict whether an NBA ‘rookie’ will have a career that spans at least 5 years. A rookie is an NBA player who is in the first year of their career.  This model has numerous applications – for example, a collectibles investor could use this prediction to inform purchases of ‘rookie’ basketball cards while they are still cheap. |
| **1.b. Hypothesis** | Using an ensemble model of different model variations will improve upon a ‘base’ random forest model.  Currently, the team’s two best performing models are a random forest and a linear regression model. Theoretically, averaging the results of the two models could improve the predictions. |
| **1.c. Experiment Objective** | I expect that using an ensemble method with different models will improve the results. Different models typically perform better in different circumstances. For example, a random forest will outperform a linear model in scenarios where the relationship between the independent variables and dependent variable are non-linear. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | A shared function `download\_data` was used to retrieve data from the Kaggle API, unzips the files and saves them into the raw data directory. This data contains two CSV files – ‘train’ and ‘test’.  The training dataset is cleaned by removing the ID columns ‘Id’ and ‘Id\_old’. The test dataset is split into ‘X\_test’ (without the id column) and ‘test\_id’ – a pandas series of the Ids.  Next, the training set is split into the independent and dependent variables using the shared function `separate\_target`.  Lastly, the training data is then split into the training and validation datasets using the ‘train\_test\_split’ function. I elected to use a test\_size of 20%. This was an arbitrary value as this aspect was not the main focus of this experiment.  There was no notable data cleaning performed on the dataset as it was a generally clean dataset with no missing values, or obvious erroneous entries. |
| **2.b. Feature Engineering** | Since this model is using the `VotingClassifier` ensemble method, different feature engineering methods were used for the two input classifiers. The input classifiers were structured as sklearn pipelines to allow for different feature inputs to the classifiers.  Both classifiers used the `StandardScaler` - which performs normal standardization:  The logistic regression model uses Principal Components Analysis to convert the independent variables to Principal Components. I did this to convert the independent variables to orthogonal variables. |
| **2.c. Modelling** | The model uses a voting classifier with a random forest classifier and a logistic regression classifier as inputs. The voting classifier uses the ‘soft’ voting method which classifies the result based on the highest average probability of the underlying classifiers.  The logistic regression has hyperparameter tuning on two hyperparameters – penalty, and PCA\_components (i.e. number of independent variables). |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. |