# **EXPERIMENT REPORT**

| **Student Name** | Tarun Gupta |
| --- | --- |
| **Project Name** | Kaggle Competition 1 |
| **Date** | 18/08/2023 |
| **Deliverables** | <Jupyter Notebook>  <Logistic Regression> |

| 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** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?  **Answer:** The business team has asked to predict the college basketball player will be drafted to join NBA league based on its statistics and previous performance metrics. The accuracy of results will help to determine the player performance and a chance to select to NBA league. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,  **Answer:** The data provided has sufficient information to analyse the insights and the information mentioned is self-explanatory to understand the requirement and its features provided. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  **Answer:** Expected outcome will be the probability of the college basketball player whether the player will be selected to join NBA league.  If the prediction comes near to 0, then the player most likely not to join the NBA league.  If the prediction comes near to 1, then the player most likely to join NBA league. |

| 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** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  **Answer:** Steps performed in preparing the data.  **Missing Data Treatment:** First step is to handle the missing data. Few of the columns having majority of data as NAN have been dropped to get the better performance of model.  **Feature Selection:** Drop the irrelevant columns which are not required for model. Also, I have performed Pearson’s correlation coefficient to select the features having lesser correlation between each other.  **Remove Outliers:** I have not performed the outliers as the data was is the different range and it was difficult to find on which feature we should perform outliers. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments  **Answer:** Steps Performed in generating features:  **Pearson’s Correlation Coefficients:** I have calculated the correlation coefficients between the input variables to understand features correlation.  **Train-Test split:** To analyse the accuracy of model, I have splitted the train data into train-test split. I have trained the model on the train data and check the accuracy on the test data. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments  **Answer:** I have chosen Logistic Regression model to perform. Logistic Regression performs well on the large features dataset. I have used cross-validation, MSE and MAE to check the accuracy of model. Also, I have used AUROC curve to validate the performance of model. |

| 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.  **Answer:** The score received from the model was quite good and the model is able to achieve the accuracy of 0.9917. |
| **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)  **Answer:** The results achieved by the model is quite satisfactory with good score value. Also, MSE and MAE is minimum(with value 0.01) |
| **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.  **Answer:** Major issue faced was in the feature selection. As, we have received 64 features dataset, it is difficult to justify which feature we need to select for modelling. I have used Pearson’s correlation coefficient and removed missing values features for the feature selection. |

| 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.  **Answer:** The results achieved by the experiment is quite good. I would like to handle missing data in more efficient way in forthcoming experiment. |
| **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.  **Answer:** This experiment gave me new insights and the projections for the next experiment like more feature selection techniques, remove outliers on multi-feature dataset. |