**KAGGLE PLAYGROUND PREDICTION COMPETITION**

# Obesity risk multiclass prediction



**Project proposal**

**MULTICLASS OBESITY RISK PREDICTION**

## **Overview**

*Development of a regularized logistic regression model and a gradient boost model for obesity risk prediciton.*

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stage** |
| 1. Data Understanding | Load datasets, explore data types, and check for missing values. | Plan |
| 2. Data Preprocessing and Exploratory data analysis | Clean data, handle missing values, and standardize features. | Analyze |
| 3. Feature Engineering | Create new features and select relevant ones for modeling. | Construct |
| 4. Model Training - OvO Logistic Regression | Train an OvO logistic regression model. | Construct |
| 5. Model Evaluation - OvO Logistic Regression | Evaluate the OvO logistic regression model using validation set. | Execute |
| 6. Model Training - LightGBM | Train a LightGBM model. | Construct |
| 7. Model Evaluation - LightGBM | Evaluate the LightGBM model using validation set. | Execute |
| 8. Model Comparison | Compare the performance of OvO logistic regression and LightGBM models. | Execute |
| 9. Model Tuning | Optimize hyperparameters for the best-performing model. | Execute |
| 10. Final Model Submission | Submit the final model to the Kaggle competition. | Execute |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

* **Audience**: Kaggle competition participants, healthcare professionals, policymakers.
* **Goal**: Predict obesity risk levels using multiclass classification.
* **Impact**: Inform public health strategies, personalized healthcare, and policy decisions.
* **Questions**: What features influence obesity risk? How well do the models generalize to unseen data?
* **Resources**: Python, scikit-learn, LightGBM, Kaggle competition datasets.
* **Deliverables**: Preprocessed datasets, feature engineering reports, model training scripts, performance metrics, final submission.

**Get Started with Python**

* **Preparation**: Familiarize with Python libraries for data manipulation and machine learning.
* **Codebooks**: Jupyter notebooks for data exploration, model training, and evaluation.
* **Activities**: Review Python basics, data science workflows, and machine learning algorithms.

**Go Beyond the Numbers: Translate Data into Insights**

* **Data Columns**: Identify relevant features for obesity risk prediction.
* **Units**: Ensure consistency in units across features.
* **Presumptions**: Anticipate potential correlations and causal relationships.
* **Missing Data**: Assess and handle missing values appropriately.
* **Format Consistency**: Ensure data is clean and in a consistent format.
* **EDA Practices**: Perform exploratory data analysis to understand data distribution and relationships.

**The Power of Statistics**

* **Purpose**: Understand the underlying distribution of obesity risk levels.
* **Research Question**: How do different demographic and lifestyle factors contribute to obesity risk?
* **Random Sampling**: Ensure the dataset is representative to avoid bias.

**Regression Analysis: Simplify Complex Data Relationships**

* **Stakeholders**: Data scientists, healthcare providers, policymakers.
* **Solution**: Develop predictive models to identify high-risk individuals.
* **Observations**: Examine correlations between features and target variable.
* **Resources**: Statistical software, domain knowledge.
* **Ethical Considerations**: Ensure privacy and confidentiality of individual data.

**The Nuts and Bolts of Machine Learning**

* **Problem**: Predict obesity risk levels based on various factors.
* **Resources**: Machine learning libraries, computational resources.
* **Data Reliability**: Assess the quality and relevance of the dataset.
* **Ethical Considerations**: Address fairness and bias in model predictions.
* **Additional Data**: Consider incorporating external data for better predictions.
* **Success Metric**: Use appropriate classification metrics (e.g., F1 score, AUC-ROC).

**Data Project Questions & Considerations**

**PACE: Analyze Stage**

* **Sufficient Information**: Evaluate if the dataset provides enough information to achieve the goal.

**Go Beyond the Numbers: Translate Data into Insights**

* **EDA Effectiveness**: Determine the most effective EDA practices for the project.
* **Data Structuring**: Decide on filtering, sorting, or other structuring methods for the dataset.
* **Visualization Assumptions**: Anticipate the types of visualizations that will be most informative.

**The Power of Statistics**

* **Descriptive Statistics**: Understand the central tendency and dispersion of obesity risk levels.
* **Null vs. Alternative Hypothesis**: Formulate hypotheses for statistical testing.

**Regression Analysis: Simplify Complex Data Relationships**

* **EDA Purpose**: Identify important features and interactions for regression models.
* **Ethical Considerations**: Consider the ethical implications of the data and model predictions.

**The Nuts and Bolts of Machine Learning**

* **Problem Revisiting**: Reassess the problem definition and model suitability.
* **Data Assumptions**: Check if the data meets the assumptions of the chosen models.
* **Variable Selection**: Justify the selection of independent variables for the model.
* **EDA Insights**: Reflect on the insights gained from EDA and how they inform the model.
* **Resources**: Utilize machine learning platforms and tools.
* **Ethical Considerations**: Address ethical considerations related to model predictions.

**PACE: Construct Stage**

* **Data Anomalies**: Identify unusual data points or averages.
* **Data Completeness**: Determine the number of data sources or groupings.

**Go Beyond the Numbers: Translate Data into Insights**

* **Data Visualizations**: Decide on the visualizations needed to communicate findings.
* **Visualization Processes**: Outline the steps to create the visualizations.
* **Variable Selection**: Choose the most relevant variables for visualizations.
* **Missing Data Handling**: Plan for handling missing data during EDA.

**The Power of Statistics**

* **Hypothesis Formulation**: Formulate the null and alternative hypotheses for testing.
* **Hypothesis Conclusion**: Draw conclusions from the hypothesis test results.

**Regression Analysis: Simplify Complex Data Relationships**

* **Model Improvement**: Evaluate the model for potential improvements.
* **Data Assumptions**: Check if the data breaks the model's assumptions.
* **Variable Selection**: Justify the selection of variables for the model.
* **EDA Insights**: Reflect on the insights from EDA and their implications for the model.
* **Resources**: Utilize statistical software and tools.
* **Ethical Considerations**: Address ethical considerations related to the model.

**The Nuts and Bolts of Machine Learning**

* **Model Problems**: Identify and address any issues with the model.
* **Model Fit**: Evaluate the model's fit to the data and its validation score.
* **Model Improvement**: Consider improvements to the model based on its performance.
* **Feature Importance**: Analyze the importance of features in the model.
* **Business Recommendations**: Propose recommendations based on the model's predictions.
* **Resources**: Utilize machine learning platforms and tools.
* **Ethical Considerations**: Ensure the model's predictions are ethical.

**PACE: Execute Stage**

* **Manager Recommendations**: Suggest further investigations based on initial data analysis.
* **Data Anomalies**: Identify anomalies in the data that may require further investigation.
* **Additional Data**: Consider what additional data could enhance the dataset.

**Go Beyond the Numbers: Translate Data into Insights**

* **Key Insights**: Summarize the insights gained from EDA and visualizations.
* **Business Recommendations**: Propose recommendations based on the insights.
* **Further Research**: Suggest additional research questions based on the findings.
* **Visualization Sharing**: Plan how to share the visualizations with different audiences.

**The Power of Statistics**

* **Business Insights**: Share key business insights from the A/B test.
* **Business Recommendations**: Propose recommendations based on the results.

**Regression Analysis: Simplify Complex Data Relationships**

* **Beta Coefficients**: Explain the importance of interpreting beta coefficients in regression models.
* **Model Recommendations**: Suggest recommendations based on the model's results.
* **Model Improvement**: Consider if the model can be improved and why.
* **Business Recommendations**: Propose recommendations based on the models.

**The Nuts and Bolts of Machine Learning**

* **Key Insights**: Share the key insights from the models.
* **Model Selection**: Explain the criteria for model selection.
* **Model Fit**: Evaluate the model's fit to the data and its acceptability.
* **Feature Importance**: Analyze the importance of features in the model.
* **Business Recommendations**: Propose recommendations based on the models.
* **Resources**: Utilize machine learning platforms and tools.
* **Ethical Considerations**: Ensure the model's predictions are ethical.

This PACE strategy provides a structured approach to your data science project, ensuring that you address all necessary steps from planning to execution.