Model Development Phase Template

Date	15 March 2024
Team ID	739849
Project Title	Doctors Annual Salary Prediction
Maximum Marks	5 Marks

Model Selection Report

In the model selection report for future deep learning and computer vision projects, various architectures, such as CNNs or RNNs, will be evaluated. Factors such as performance, complexity, and computational requirements will be considered to determine the most suitable model for the task at hand.

Model Selection Report:

Model	Description
Linear Regression	This is a fundamental machine learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It's simple and easy to interpret. It assumes a linear relationship between the dependent and independent variables. While it is computationally efficient, it may not capture complex patterns in the data.
Random Forest	An ensemble learning method that constructs multiple decision trees and merges them to get a more accurate and stable prediction. It reduces overfitting, handles large datasets with higher dimensionality, and improves accuracy by averaging the results of many decision trees. However, it requires more computational resources and memory.

Decision Tree	A non-linear model that splits the data into subsets based on the value of input features. It is easy to visualize and understand, handles both numerical and categorical data, and can capture non-linear relationships. However, decision trees can be prone to overfitting, especially with deep trees.
Gradient Boosting	An ensemble technique that builds models sequentially. Each model attempts to correct the errors of the previous one. It is highly effective and provides high predictive accuracy. GBMs can handle various types of data and are less prone to overfitting compared to single decision trees. However, they are more complex and computationally intensive.
Support Vector Machine (SVM)	A powerful classifier that works by finding the hyperplane that best separates the classes in the feature space. It is effective in high-dimensional spaces and robust to overfitting, especially in cases where the number of dimensions exceeds the number of samples. SVMs require careful tuning of parameters and are computationally intensive, especially with large datasets.
K-Nearest Neighbors (KNN)	A simple, instance-based learning algorithm where the prediction for a new data point is made based on the majority class among its k-nearest neighbors. It is easy to implement and understand but can be computationally expensive as it stores all the training data and requires efficient techniques for finding nearest neighbors in large datasets.
Neural Networks	A set of algorithms modeled loosely after the human brain, designed to recognize patterns. They consist of layers of interconnected nodes where each layer transforms the input data into increasingly abstract representations. Neural networks can model complex relationships and capture intricate patterns in the data. However, they require large amounts of data and computational power, and the training process can be time-consuming.
XGBoost	An optimized implementation of gradient boosting designed to be highly efficient, flexible, and portable. It provides parallel tree boosting to solve many data science problems quickly and accurately. XGBoost is known for its speed and performance and can handle large datasets efficiently.

	However, it requires careful parameter tuning to achieve optimal performance.