COMP9321 Data Services Engineering

Plan of assignment 3

Group name: Macau casino

Group member: Jiachen Li, Patrick Wang, Lesi Huang, Xiyao Li, Chong Zu

Introduction

Aim: In this assignment, we are going to establish a data service to analyze factors which have been effecting Australian GDP during the past ten years, and predict the trend of the GDP according to these factors in next 1-5 years.

Datasets: Kaggle Datasets
(https://www.kaggle.com/datasets)
NSW Government Public Datasets
(https://data.nsw.gov.au/data/dataset)
Australian Federal Government Public Datasets.
(https://search.data.gov.au/)

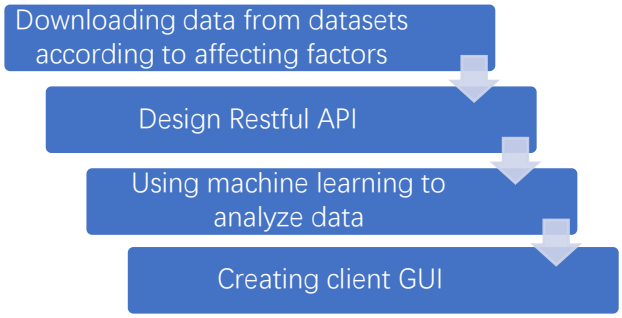
Communication channel: group chat in Wechat.

Code repository: https://github.com/CocaColaSoap/australia-GDP-analysis

Task

Members role: Our team would be divided into two sub-teams. The first team (Jiachen Li and Chong Zu) would be responsible for Restful API and GUI Design. The second group (Patrick Wang, Lesi Huang, Xiyao Li) would work on machine learning model to integrate, process and analyze data collection.

Work flow:



Machine learning model selection:

At this stage, we are considering three machine learning models .

1. First model is **linear regression based method**. The basic assumption is that the output variable (a numeric value) can be expressed as a linear combination (weighted sum) of a set of input variable (which is also numeric value).

y = w1x1 + w2x2 + w3x3 ....

The strength of Linear model is that it has very high performance in both scoring and learning. The Stochastic gradient descent-based learning algorithm is highly scalable and can handle incremental learning.

The weakness of linear model is linear assumption of input features, which is often false. Therefore, an important feature engineering effort is required to transform each input feature, which usually involved domain expert.

2. We also consider to use **Neural Network model**. Neural Network can be considered as multiple layer of perceptron (each is a logistic regression unit with multiple binary input and one binary output). By having multiple layers, this is equivalent to : z = logit(v1.y1 + v2y2 + ...), while y1 = logit(w11x1 + w12x2 + ...)

This multi-layer model enables Neural Network to learn non-linear relationship between input x and output z. The typical learning technique is "backward error propagation" where the error is propagate from the output layer back to the input layer to adjust the weight.

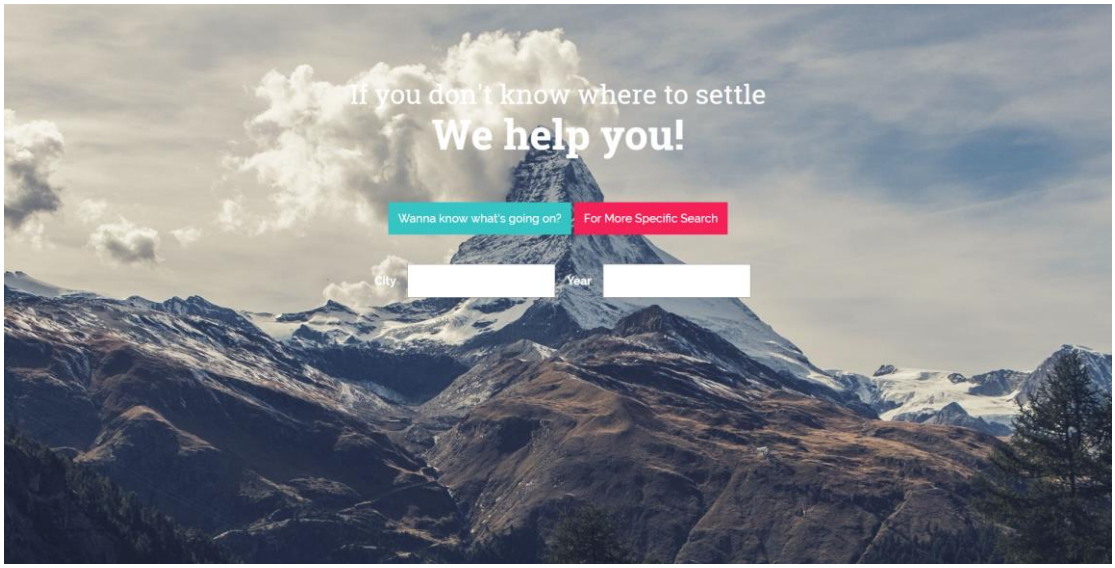
Notice that Neural Network expect binary input which means we need to transform categorical input into multiple binary variable. For numeric input variable, we can transform that into binary encoded 101010 string. Categorical and numeric output can be transformed in a similar way.

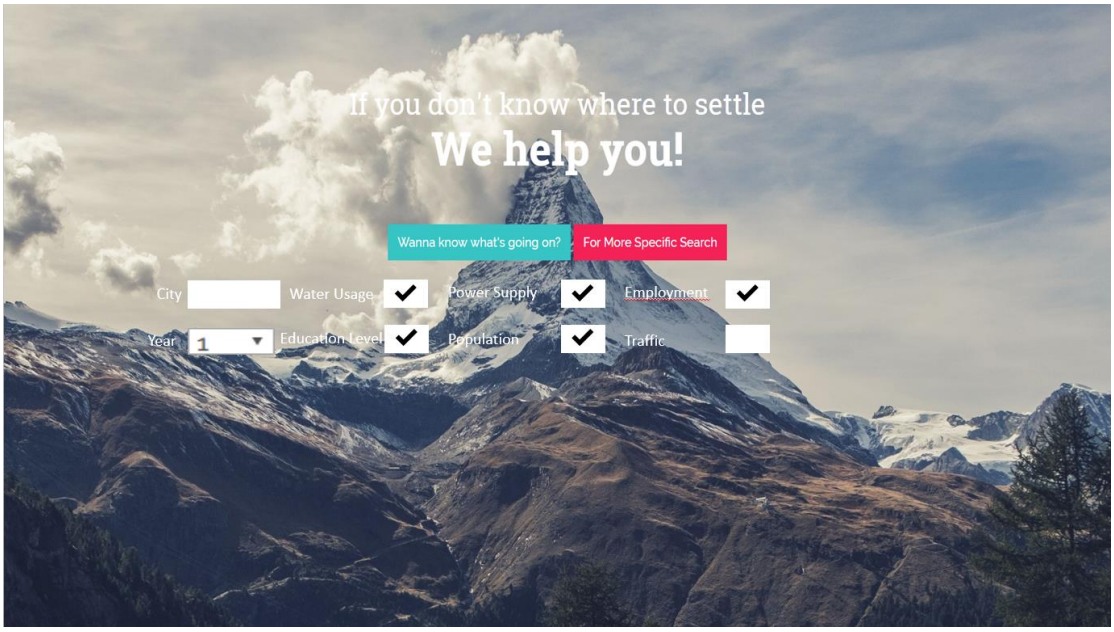
3. The last model we want to try is **Nearest Neighbor model**.

The idea is to find K similar data point from the training set and use them to interpolate the output value, which is either the majority value for categorical output, or average (or weighted average) for numeric output. K is a tunable parameter which needs to be cross-validated to pick the best value.

The strength of K nearest neighbor is its simplicity as no model needs to be trained. Incremental learning is automatic when more data arrives (and old data can be deleted as well). Data, however, needs to be organized in a distance-aware tree such that finding the nearest neighbor is O(logN) rather than O(N). On the other hand, the weakness of KNN is it doesn't handle high number of dimensions well. Also, the weighting of different factors needs to be hand tuned (by cross-validation on different weighting combination) and can be a very tedious process.

Interface design:





At this stage, we design an interface which is very convenient for users to input different affecting factors in order to get the statistics of GDP in past ten years of get the prediction of GDP in next 1-5 years.

For the most important filter conditions, two searching blanks are designed which are “City” and “Year” respectively. Each searching blank provides many options or different selection interval.

As the output part, the statistics during the past period or predicted in the future are expected to be shown as graphs such as bar chart and line chart, which are straightforward for trends demonstration.