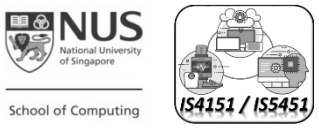

IS4151/IS5451 – AIoT Solutions and Development

AY 2024/25 Semester 2

Practical Lab 07 – Machine Learning for IoT Data (I)

Part 1 – Basic Programming

Point prediction of marketvalue – The following series of exercises is based on the Forbes2000.csv dataset and you will be building some simple models based on marketvalue. In particular, perform the tasks in each of the following exercises and report your results and observations.

PE07-1-1 – Simple Linear Regression

Build three different and separate simple linear regression model to predict marketvalue using sales, profits and assets. You should use both Scikit Learn (https://scikit-learn.org/stable/) and statsmodels (http://www.statsmodels.org/stable/index.html). statsmodels can be installed with the following command using pip:

```
python -m pip install statsmodels
```

Which model, or independent variable, is the best one for predicting marketvalue?

PE07-1-2 – Multiple Linear Regression

Point prediction of marketvalue – Build a single multiple linear regression model to predict marketvalue using sales, profits and assets. You should use both Scikit Learn and statsmodels.

How does this multiple linear regression model compare with the simple linear regression models in PE07-1-1?

PE07-1-3 – Multiple Linear Regression with Categorical Independent Variables

Point prediction of marketvalue – Build a multiple linear regression model to predict marketvalue using sales, profits, assets and category.

Category is a categorical independent variable. Perform dummy encoding by omitting one specific category. Choose the most appropriate category to be omitted. Which category did you omit and why?

Use both Scikit Learn and statsmodels to fit the model. How does this multiple linear regression model compare with the one in PE07-1-2?

PE07-1-4 – Classification

Binary classification of marketvalue (low and high) – Build a decision tree classifier to predict marketvalue as a binary variable (see PE06-2-1-g) using sales, profits and assets. You should use Scikit Learn.

We would be using Graphviz to generate the decision tree visually. Graphviz can be installed with the following command using pip:

```
python -m pip install graphviz
```

You also need to download the Graphviz executable package from here – https://graphviz.org/download/

How does this classification model compare with the multiple linear regression models in PE07-1-2?

Part 2 – Advanced Programming

PE07-2-1 – Predict House Price with Linear Regression

This exercise is based on a dataset on house sales in King County, USA. The dataset is taken from Kaggle – https://www.kaggle.com/harlfoxem/housesalesprediction.

The dataset consists of 21,613 observations of recently transacted properties, and contains the following seventeen variables:

- id – A unique numerical identifier for a house.
- date – Date the house was sold.
- price – Selling price of the house, this is the **dependent or target variable**.
- bedrooms – Number of bedrooms/house.
- bathrooms – Number of bathrooms/house.
- sqft_living – Square footage of the house.
- sqft_lot – Square footage of the lot.
- floors – Total floors (levels) in the house.
- waterfront – House that has a view to a waterfront.
- view – How many times the house has been viewed.
- condition – How good the overall condition of the house is.
- grade – Overall grade given to the housing unit, based on a standardised grading system.
- sqft_above – Square footage of house apart from basement.
- sqft_basement – Square footage of the basement.
- yr_built – Built year.
- yr_renovated – Year when house was renovated.
- zipcode – Zip or postal code of the house.

a) Perform an exploratory data analysis on the dataset and report your results and findings.

b) Perform linear regression analysis on the dataset and report your results and findings:

   i. What are some simple and/or multiple linear regression models that you have trained?

   ii. Which one of these models is the best model for predicting the selling prices of houses?

PE07-2-2 – Predict Adult Census Income with Classification

This question is based on the Census Income (Adult) dataset taken from the UCI Machine Learning Repository – http://archive.ics.uci.edu/ml/datasets/Adult.

The original source of the dataset is attributed to:

Kohavi, R. and Becker, B., Data Mining and Visualization, Silicon Graphics.

The dataset was extracted from the 1994 United States Census Bureau database (originally found at http://www.census.gov/ftp/pub/DES/www/welcome.html). The filtering criteria were "((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))". The prediction task is to determine **whether a person makes over $50000 (i.e., 50K) a year**.

The dataset downloaded from UCI Machine Learning Repository consists of more than 30000 observations. There is a target variable together with 14 other variables. The actual dataset given to you has the variable fnlwgt (final weight) removed. Description of the target variable and the remaining 13 predictive variables are listed below:

1. Income – <=50K (coded as lte50k) or >50K (coded as gt50k)
2. age – continuous.
3. workclass – Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
4. education – Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: Number of years of education – continuous.
6. marital-status – Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation – Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Relationship of individuals to the head of household – Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race – White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex – Female, Male.
11. capital-gain – continuous.
12. capital-loss – continuous.
13. hours-per-week: Hours worked per week – continuous.
14. native-country – United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

a) Perform any data preparation that is required together with an exploratory data analysis on the dataset and report your results and findings.

b) Perform a decision tree analysis and report your results and findings.