

***For the module MATH9001 – Research Methods***

***Master of Science in Data Science and Analytics, Department of Mathematics, date***

**Declaration of Authorship**

I, Vaishnavi Kale , declare that this thesis titled ‘ Regression and Feature Selection on the Family Income and Expenditure Survey ’ and the work presented in it are my own. I confirm that,

* This work was done wholly or mainly while in candidature for the Masters’ degree at Cork Institute of Technology
* Where any part of this thesis has previously been submitted for a degree or any other qualification at Cork Institute of Technology or any other institution, this has been clearly stated
* Where I have consulted the published work of others, this is always clearly attributed
* Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work
* I have acknowledged all main sources of help
* I understand that my project documentation may be stored in the library at CIT, and may be referenced by others in the future

Signed:

Date:

Table of Contents

[Abstract 4](#_Toc35941557)

[1. Introduction 3](#_Toc35941558)

[2. Research 5](#_Toc35941561)

[3. Methodology 7](#_Toc35941562)

[4. Evalution 9](#_Toc35941563)

[5. Conclusion and Future Work 14](#_Toc35941564)

[References 15](#_Toc35941564)

# 

# Abstract

Socio Economic Classification plays a vital role in measuring individual’s economic and sociological standing. Hence, it is important to have a widely acceptable model that is useful for the government as well as researchers of any country. In this paper, The Family Income and Expenditure Survey (FIES) data of Philippines is analysed to find the best model to predict household income and selecting key features that affect income of families in Philippines. First, feature selection is using following approaches : (i) Univariate Feature Selection (ii) Tree-based Feature Selection, and (iii) Greedy Feature Selection with Recursive Elimination approach. Second, prediction task is accomplished by following models: (i) Multiple Linear Regression (ii) k- Nearest Neighbors, and (iii) Random Forest for Regression. Using decision tree and linear regression estimator with RFE feature selection method, the number of important variables found are 26 and 46 respectively. The error rate for models with 26 and 46 variables are almost equal. The results are competitive for models such as linear regression and random forest with root mean square error as 0.87% and 0.89% respectively. Our best model is linear regression with selected 26 variables which produces mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and coefficient of determination (R2 Score) of 0.44%, 0.01%, 0.87%, 87.94% respectively.

# 

# 1. Introduction

Every three years, the survey is undertaken by the Philippine Statistics Authority (PSA) spearheads to provide data on family income, expenditure on different items, sources of income in cash, and related information affecting income and expenditure levels and patterns in the Philippines. The Family Income and Expenditure Survey (FIES) is carried out nationwide. The household income is the biggest indicator of socio-economic classification in any country. However, the socio-economic classification models in the Philippines are highly questionable. In fact, not one SEC model has been widely accepted. Government bodies uses their own SEC models and private research entities uses their own. Therefore, the aim of the paper is to find the best model for prediction of household income and some key drivers of household income.

There is one research work done on Income and Expenditure in 2009. In this paper [1], the author predicted fiscal income and expenditure of Guangfeng county from 2001 to 2006 with the grey model. Another work related to this is research on balance between income and expenditure in paper [2]. The aim of this paper was to develop the predictive model of balance between income and expenditure of medical insurance fund. Furthermore, the specific policy was suggested.

The data set used in this paper is provided by the Philippine Statistics Authority. It has some selected variables from the latest Family Income and Expenditure Survey (FIES) in the Philippines. It contains more than 40,000 observations and 60 features which primarily consist of the household income, expenditures, personal and family information of that household. Our target regression variable that we want to predict is household income using the most efficient model.

As discussed earlier, two main objectives of this paper are feature selection and choosing the best model for prediction of the income value. The following three approaches to accomplish the task of feature selection:

1. Univariate Feature Selection
2. Tree-Based Feature Selection
3. Greedy backward Feature Selection

The feature selection algorithms are discussed in Section 2 (named as Research). Next, Section 3 has methodology which compares several machine learning algorithms with detail description. Following are the three methods used for regression modelling:

1. Multiple Linear Regression
2. k Nearest Neighbours Regression Algorithm
3. Random Forest.

The results of the feature selection and models are examined in Section 4 (Evaluation). Finally, conclusion and possible future works in this area or dataset is in Section 5.

# 2. Research

# The dataset contains various parameters such as expenditure spent on different item, number of appliances used and members in the family, etc. The topic of research is feature selection as will not only help finding key factors affecting household income but also make regression models sparse and efficient. Following are the three methods used for this dataset:

# 2.1 Univariate Feature Selection:

# Univariate feature selection evaluates the relationship of the feature with the response variable individually and calculates the strength correlation which is given as score of that variable. There are many univariate methods for feature selection which are simple and good for understanding of data. However, not necessarily helpful for optimizing the feature set for better generalization. One of popular SelectKBest function where K is number of highest scores is used for feature selection in this dataset. For regression models, sklearn provides f\_regression method which computes the p-values and F-statistics for feature selection. The k value is set to 20. Results are discussed in Evaluation Section.

# 2.2 Tree-based method:

# Decision trees is classification or regression process in which a tree is built where internal nodes are like if else statement and leaf nodes their outcomes. It reduces the uncertainty or noise in data. For regression the measure of impurity is variance. While training tree it is possible to calculate the amount of reduction in impurity caused by a feature. The variable with highest decrease in impurity is most important feature. Rest of features can be ranked according to their importance. Here, random forest method is applied. Two main reason for using Random forest over decision tree are:

# Random Forest is ensemble learning technique which is based on bagging algorithm. It builds as many trees on random subsets of the data and combines the output of all trees. As a result, it reduces overfitting problem in decision tree as well as reduces the variance in the data.

# And most importantly, provide a reliable feature importance estimate as it averages the reduction in variance (noise) achieved by each feature across all trees generated.

# 2.3 Greedy Feature Selection:

# It is a greedy optimization algorithm like recursive feature elimination (RFE) which aims to find the best performing feature subset. It chooses features by recursively makes smaller and smaller set of features. Following is the working of RFE:

# Initially all the features are included in a feature set and the importance or score of each feature is calculated as the problem is regression simple linear regression model coefficients are used.

# The feature with least importance is removed from the current feature set.

# This process is repeated recursively until all the features are ranked.

# This method is important as it gives the number of important variables in the whole dataset. In this work, both decision tree and linear regression methods are used as RFE estimator.

# 3. Methodology

**3.1 Data Pre-processing :**

1. Exploratory data analysis is carried out on outcome variable and some features to check the distribution of income (target variable) and relationship with other variables.
2. Two categorical features in the dataset have missing values. Since the missing values for both the columns are less than 20%, these variables are imputed by statistical method called as mode or most frequent value of the column.
3. There are 15 categorical variables in this dataset. As these features are nominal values that is, they have no inherent ordering, Ordinal Encoder method from sklearn package is used without need of explicitly mapping the features. Additionally, unlike one-hot-encoding method, it does not increase the dimensionality by creating dummy variables.
4. Next step is feature scaling. The features in the dataset are normalized into range between 0 and 1. Because, normalization can improve the performance of algorithms such as linear regression which assign weight to features and k-nearest neighbours’ regression which calculates geometric distance.

**3.2 Multiple Linear Regression :**

Multivariate linear regression is a generalization of linear regression by considering more than one independent variable. The dependent (target) variable *Y* has a linear relationship with the independent variables *x*1,*x*2,⋯,*xn*, the basic model for linear regression is [3]:

*Y* = *β*0 + *β*1*x*1 + ⋯+ *βn xn* + *ε* (1)

where *β*1,*β*2,⋯,*βn* are the regression coefficients ; E is a random item, *E*(*ε*)=0.

**3.3 k-Nearest Neighbors Regression:**

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

1. Choose the value of K i.e. the nearest data points.
2. Calculate the distance between test data and each row of training data with the help of Euclidean distance.
3. Sort them in ascending order.
4. Choose the top K rows from the sorted array.
5. The predicted value is average of the top k point from sorted list.

After trying several values for k, the optimal value for k is 6 for this dataset.

**3.3 Random Forest:**

A Random Forest is an ensemble technique which creates multiple decision trees on sub sampled training data. The final output is determined by the result of multiple decision trees the rather than relying on individual decision trees.

**Algorithm:**

1. Select k number of trees
2. Resample the training set into N samples.
3. Fit the decision tree on N samples.
4. Repeat steps 2 and 3
5. For test prediction , predict the value of target variable with k trees and the final result is average across all the predicted values by k trees.

Number of trees (n\_estimators) selected is 250 as after 250 trees accuracy of model either decreases or remained constant.

The dataset is split into train (70%) and test (30%) sets. 10-fold cross validation with scoring parameter as root mean square error is used. Above mentioned models are applied to 26 selected features using decision tree as well as 46 variables subset selected using linear regression estimator in RFE with 10-fold cross validation. The performance of all these models are evaluated in next section.

# 

# 4. Evaluation

# First step is evaluation of feature selection methods:

# 4.1 Univariate Feature Selection Result:

|  |  |  |
| --- | --- | --- |
| Sr No | Features | Scores |
| 1 | Communication Expenditure | 42262.55393 |
| 2 | Housing and water Expenditure | 34994.642705 |
| 3 | Miscellaneous Goods and Services Expenditure | 33165.794473 |
| 4 | Total Food Expenditure | 32699.075869 |
| 5 | Transportation Expenditure | 28811.131661 |
| 6 | Clothing, Footwear and Other Wear Expenditure | 24108.582187 |
| 7 | Imputed House Rental Value | 20422.171591 |
| 8 | Meat Expenditure | 19697.797600 |
| 9 | Total Income from Entrepreneurial Activities | 19344.552663 |
| 10 | Number of Personal Computer | 18740.182133 |

# Table-1 : Top-ten highest scored variables

# 4.2 Tree-based Feature Selection Result:

# C:\Users\Vaishnavi\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\8743A320.tmp

# Figure-1: Bar graph of important variables

# Ranking of variables by pre-process statistical univariate method and tree-based method is different. Highest scored variable for SelectKBest function is communication expenditure whereas for random forest it is Miscellaneous Goods and Services Expenditure. From both the methods, amount of money spent on Communication, Housing and water, Food, Transportation, Miscellaneous Goods and Services , Clothing , Footwear and other wear are some important parameters for predicting household income value.

# 4.3 RFE Result:

# Using linear regression estimator:

# C:\Users\Vaishnavi\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5779FDC2.tmp

# Figure-2 : Line graph of variables for decision tree estimator

# The score of the variables remains constant after 40 variables. Hence, it gives 46 important variables. Then, models mention in section-3 are applied on the subset of these variables.

# Using decision tree estimator:

# C:\Users\Vaishnavi\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\4631E3C7.tmp

# Figure-3 : Line graph of variables for linear regression estimator

# The score of the variables has slight drop at approx. 5 and then it increases till 20. After 20 variables, there is more fluctuations and score stayed constant. Hence, it gives 26 important variables. Then, models mention in section-3 are applied on the subset of these variables.

# 4.4 Comparing performance of the models:

# Following are the metrics for assessing performance of models:

# Mean Square Error (MSE):

# MSE is the average of the squared difference between the target value and the value predicted by the regression model. As the differences are squared, it penalizes even a small error which leads to over-estimation of the performance of model. It is preferred more than other metrics because it is differentiable and hence can be optimized better.

# MSE =

where, N is number of observations

y = target value

= predicted value by the model

# Mean Absolute Error (MAE):

# The absolute difference between the target value and the value predicted by the model is given by MAE. In this, all the individual differences are weighted equally it is a linear score. The MAE does not penalize the errors as extremely as mean square error and also it is more robust to outliers.

# MAE =

where, N is number of observations

y = target value

= predicted value by the model

# Root Mean Square Error (RMSE):

# Root Mean Square Error is the square root of the averaged squared difference between the target value and the value predicted by the model. It is a high penalty on large errors because the errors are first squared before averaging them. This implies that RMSE is useful when large errors are undesired.

# RMSE =

where, N is number of observations

y = target value

= predicted value by the model

# R-Squared Score (R2) :

R squared also known as coefficient of determination is the proportion of the variance in the dependent variable that is predictable by all the independent variables together. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

# R2 = 1 -

# Comparing performance of the models with selected 46 variables using RFE with linear regression estimator:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MSE | RMSE | R-square |
| Linear Regression | 0.0044 | 0.0001 | 0.0087 | 0.8496 |
| k-NN Regression | 0.0061 | 0.0002 | 0.0125 | 0.6862 |
| Random Forest | 0.0041 | 0.0001 | 0.0089 | 0.8419 |

# Table-3 : Performance table of selected variables

Linear regression and random forest both models give good result as they have below 0.9 % error rate that is root mean square error. About 84% of total variation in target variable is explained by these models. However, performance of k nearest neighbors is low as compared to other two models as the error (all three i.e. MAE, MSE, RMSE) is high for k-NN model. Additionally, the variance in outcome variable explained by k-NN is just 68.62%. The result of 10-fold cross validation is given below:

# C:\Users\Vaishnavi\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\35746203.tmp

# Table-2 : Performance table of selected model

# Comparing performance of the models with selected variables using RFE with decision tree estimator in which has subset of 26 explanatory variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MSE | RMSE | R-square |
| Linear Regression | 0.0044 | 0.0001 | 0.0087 | 0.8497 |
| k-NN Regression | 0.0064 | 0.0002 | 0.013 | 0.6597 |
| Random Forest | 0.0041 | 0.0001 | 0.0089 | 0.8431 |

# Table-2 : Performance table of model with all variables

# Using 26 variables gives almost same result as that of models with 46 expalanatory variables. All the error namely MAE, MSE and RMSE measured are nearly equal. However, the coefficient of determination which is given by R-squared score is slightly decresed. To conclude, 26 variables are enough

# for prediction of target variable (hosehold income). Below given is result of 10-fold cross validtion for three models:

# C:\Users\Vaishnavi\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F96C486D.tmp

# Figure-3 : Boxplot of all the models with all variables

# Linear model is best for this dataset as its median value is slightly higher than random forest. Furthermore, the interquartile and actual range for linear regression boxplot is smaller compared to other models. This means the error rate for 10-fold cross validation models using linear regression has error rate between 0.08 to 0.1 with 2 exceptions (outliers).

# 5. Conclusion and Future Work

In this work, three different approaches are presented for feature selection as follow: (i) Univariate Feature Selection (ii) Tree-based Feature Selection, and (iii) Greedy Feature Selection with Recursive Elimination approach. For prediction model methods are : (i) Multiple Linear Regression (ii) k- Nearest Neighbors, and (iii) Random Forest for Regression. Initially I applied feature selection in which I found some expenditure variables such as communication, Miscellaneous Goods and services , etc as important variables in deciding income value. Using different feature selection estimators in RFE set of 26 and 46 variables are selected. I then applied three above mentioned models on subsets of selected variables. The best model is linear regression with 26 variables with RMSE of 0.87% and 84.97% total variance in target variable (household income) is explained by the model. The future plan is to apply other regression estimators for feature selection process. I will also use other efﬁcient machine learning techniques in combination of different features selected.

# 

# References

[1] C. Shu, “Grey model of county-level fiscal income expenditure and application,” in *2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009)*, Nov. 2009, pp. 678–681, doi: 10.1109/GSIS.2009.5408227.

[2] H. Zhang and K. Wang, “Research on Balance between Income and Expenditure in the Mode of Overall Covering by Medical Insurance in Dalian,” in *2010 International Conference on E-Product E-Service and E-Entertainment*, Nov. 2010, pp. 1–4, doi: 10.1109/ICEEE.2010.5660198.

[3] Y. Yang, “Prediction and analysis of aero-material consumption based on multivariate linear regression model,” in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Apr. 2018, pp. 628–632, doi: 10.1109/ICCCBDA.2018.8386591.