**NAME: WAIRIUKO SAMUEL**

**LINEAR REGRESSION, KNN, ADABOOST, RANDOM FOREST, SVR**

**DATASET TITLE: SEATTLE HOUSE SALE**

**1.0 BACKGROUND**

This project is about building a prediction model to predict house prices I got the dataset from [www.kaggle.com](http://www.kaggle.com). The dataset had 21 variables and 21612 observations.

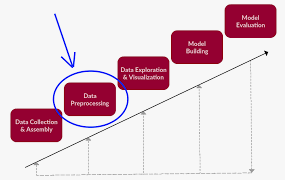
**PROPERTIES OF DATASET;**

Variables include the target variable Price, other variables include id, date, bedroom, bathrooms, square foot living, square foot lot, floors, waterfront, view, condition, grade,

square foot above, square foot basement, year built, renovated, zip code

Lat, long, sqft\_living15, sqft\_lot15,

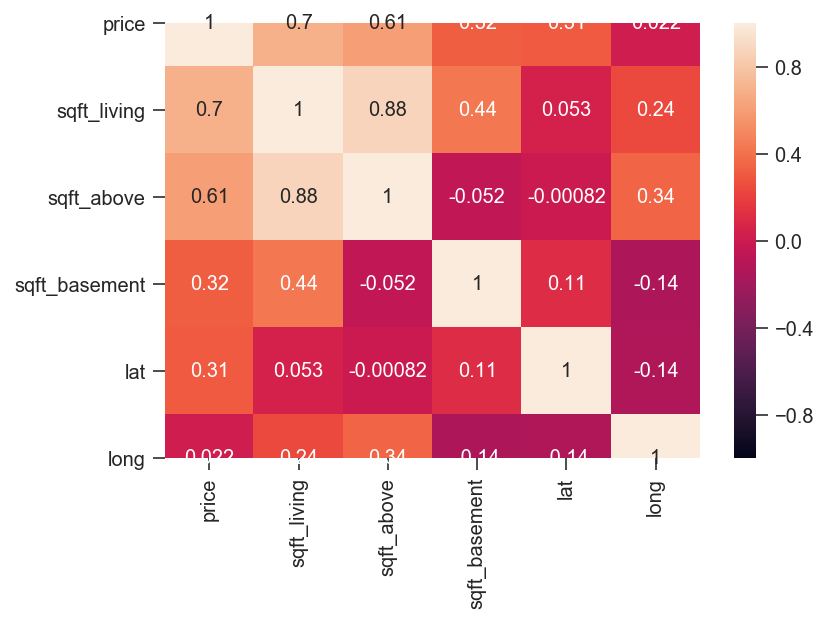
**2.0 DATA PREPARATION, CLEANING & ANALYSIS**



* Checked and removed duplicates
* Standardized the dataset
* Dropped irrelevant columns. data=dataset.drop(['id','date','zipcode','waterfront','grade','view','condition','sqft\_living15','floors','bathrooms','yr\_renovated','sqft\_lot','sqft\_lot15','yr\_built','bedrooms'],axis=1)

**3.0 DATA ANALYSIS & SUMMARY FINDINGS**

CORRELATION MATRIX



PVALUES

coef std err t P>|t| [0.025 0.975]

---------------------------------------------------------------------------------

sqft\_living 182.5811 1.765 103.436 0.000 179.121 186.04

sqft\_above 78.3612 2.167 36.162 0.000 74.114 82.60

sqft\_basement 104.2199 3.106 33.552 0.000 98.131 110.3

lat 5.513e+05 1.44e+04 38.418 0.000 5.23e+05 5.79e+05

long 2.146e+05 5579.259 38.467 0.000 2.04e+05 2.26e+

**3.1 MODEL PERFORMANCE COMPARISONS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LINEAR REGRESSION | | | | | |
|  | LINEAR REGRESSION | KNN | ADABOOST | RANDOM FOREST | SVR |
| K. FOLD | 0.58 | 0.0069 | 1 | 0.79 | 0.46 |
| R2 | 0.58 | Computer memory problem | Computer memory problem | 0.78 | -.04 |
| RMSE | 242150 | Computer memory problem | Computer memory problem | 163799 | 378277 |
| P. VALUES | 0 | 0 | 0 |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**3.2 R2, RMSE,** K. FOLD

Random forest had the highest R2 of o.78 meaning there is variability of the response data around its mean. while SVR had -0.4 meaning there is no variability.

RMSE with the lowest score was for random forest with $163799 compared to the highest $378277 from SVR.so the root mean square error for random forest is the lowest means a better model compared to the other models

**4.0. CONCLUSIONS AND RECOMMENDATIONS**

Comparing the R2, K fold and RMSE, Random forest was found to be the best model for house price prediction as compared to the other models.

**. 5.REFFERENCES**

-www.kaggle.com {open data set}

-Python Project Codes(attached)

**TITLE: CLASSIFICATION USING LOGISTIC REGRESSION, KNN, ADABOOST, RANDOM FOREST, SVM**

**TITTLE: TURKIYE STUDENT EVALUATION**

1. **INTRODUCTION**

This project is about building a prediction model to predict student grades. I got the dataset from [www.kaggle.com](http://www.kaggle.com) and analyzed it using Logistic regression, KNN, Adaboost, Random Forest and SVM models.

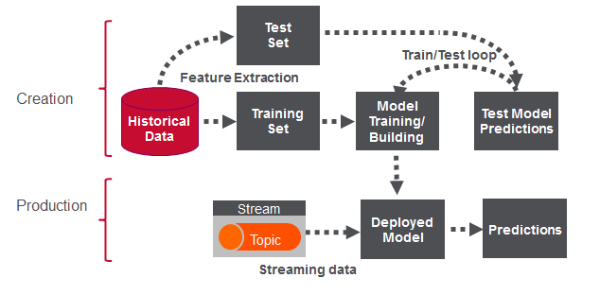
1.1 DATASET DESCRIPTION

The dataset had 28 variables and 5820 observations. The target variable was Q28.other variables were; difficulty,instr,class,nb.repeat,attendance,Q1,Q2,Q3,Q4,Q5,Q6,Q7,Q8,Q9,Q10,Q11,Q12,Q13,Q14,Q15,Q16,Q17,Q18,Q19,Q20,Q21,Q22,Q23,Q24,Q24,Q26,Q27,Q28.

**1.2 OBJECTIVES**

* To build precise models for predicting student performance.
* To compare the model performance and results.
* To analyze the dataset and give constructive feedback to the school management.

**2.0 DATA PREPARATION, CLEANING & ANALYSIS**

[](https://www.google.ca/url?sa=i&url=https%3A%2F%2Fmapr.com%2Fblog%2Freal-time-credit-card-fraud-detection-apache-spark-and-event-streaming%2F&psig=AOvVaw2ghL1uMDaSbPIwXvJi_6jG&ust=1582556729755000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPC02cb55-cCFQAAAAAdAAAAABAN)

I followed the following steps;

* Imported the Libraries
* Imported the dataset
* The data was checked for duplicates and missing values
* irrelevant variables dropped (difficulty, instr, class, nb. repeat, attendance).
* Selected the appropriate model and import from sklearn
* Divided the dataset into training and set.
* Trained the model on training set
* Evaluated the model on the test set

**3.0 DATA ANALYSIS & SUMMARY FINDINGS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CLASSIFICATION | | | | | |
|  | LOGISTIC REG | KNN | ADABOOST | RANDOM FOREST | SVM |
| K. FOLD | 0.88 | 0.87 | 0.98 | 0.99 | 1 |
| ACCURACY | 0.87 | 0.86 | 0.98 | 0.99 | 1 |
| CONFUSION MATRIX |  |  |  |  |  |
| MACRO RECALL | 0.87 | 0.86 | 1 | 0.99 | 1 |
| MACRO PRECISION | 0.90 | 0.87 | 1 | 0.99 | 1 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CLASSIFICATION AFTER SVC GRID SEARCH | | | | | |
|  | LOGISTIC REG | KNN | ADABOOST | RANDOM FOREST | SVM |
| linear | 1 | 1 | 1 | 1 | 1 |
| rbf | 0.98 | 0.99 | 0.98 | 0.98 | 1 |
| poly | 0.99 | 0.88 | 0.86 | 0.99 | 1 |

Adaboost confusion matrix

[ 67, 45, 116, 111, 82],

[ 32, 33, 78, 90, 63],

[118, 67, 215, 229, 191],

[111, 73, 238, 252, 165],

[ 63, 46, 145, 174, 106]

Logistic regression confusion matrix

[[250 0 0 0 0]

[ 0 13 162 0 0]

[ 0 0 448 41 0]

[ 0 0 20 493 0]

[ 0 0 0 0 319]]

Random forest confusion matrix

[209, 2, 1, 0, 0],

[ 4, 141, 1, 2, 0],

[ 0, 0, 411, 0, 0],

[ 0, 1, 1, 414, 0],

[ 0, 0, 0, 0, 268]

SVM confusion matrix

[212, 0, 0, 0, 0],

[ 0, 148, 0, 0, 0],

[ 0, 0, 411, 0, 0],

[ 0, 0, 0, 416, 0],

[ 0, 0, 0, 0, 268]

**4.0. CONCLUSIONS & RECOMMENDATIONS**

**4.1 RECOMMENDATIONS**

Before the grid search SVM was the best with 100% followed by Random Forest (0.99),followed by Adaboost (0.98), followed by logistic (0.88) and last KNN (0.87)

After CVC grid search all the models increased accuracy depending on the kennel used, linear kennel produced the best score as compared to rbf and poly

* 1. **CONCLUSIONS**
* All the models after grid search produced good results with 100% accuracy. The company should go for the model that is cost effective amongst the five models.

**5.REFFERENCES**

* www.kaggle.com {open data set}
* Python Project Codes(attached)