MACHINE LEARNING_PROJECT_ROBSON_MUWANI

PROJECT_A_REGRESSION_PREDICTION _MODELS

BACKGROUND SUMMARY

This report provides supervised machine learning task based on **REGRESSION MODELS**. The aim is to predict appliance energy use based on features like temperature, humidity, windspeed and dew point. The data set predictors: Temperature, Humidity, Wind Speed, Dew point were recorded in different room set up. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non predictive attributes (parameters).

REGRESSION DATASET

Sr No	Dataset Name	Description	Rows	Columns
1	**Appliance Energy Prediction**	Based on Predictors: Temperature, Humidity, Wind Speed, Dew Point we are predicting Appliance Energy use in (kWh)	**19,375**	**29**
2	**Attribute Information**	**Based on Predictors like:** T1, Temperature in kitchen area, in Celsius RH_1, Humidity in kitchen area, in % T2, Temperature in living room area, in Celsius RH_2, Humidity in living room area, in % T3, Temperature in laundry room area RH_3, Humidity in laundry room area, in % T4, Temperature in office room, in Celsius RH_4, Humidity in office room, in % T5, Temperature in bathroom, in Celsius RH_5, Humidity in bathroom, in % **etc.**	**19,375**	**29**

From the dataset the two random variables (RV1, RV2) that were included were later on dropped together with the Visibility feature as the correlation to appliance energy used was negligible, very weak and close to zero.

CORRELATION ANALYSIS

Correlation analysis from stats model reported r-square of 0.152 meaning that 15.2% of the variation in appliance energy use was explained by the predictors like temperature, humidity, and wind speed in different room set ups as stated above. After feature pruning, the model variables were reduced from 26 to 19 columns that were fed into the final model.

MODEL RESULTS SUMMARY

REGRESSION						
Energy data set	**Cross Validation Method**	**LR**	**KNN**	**Random Forest**	**SVR**	**ADABOOST**
	R_Square(r2)	**0.1294**	**0.4760**	**0.634**	**0.00084**	**-4.178**
	K-Fold Score	**0.0995**	**0.4316**	**0.6014**	**0.0077**	**-5.4061**
	K-Fold s.t.d.v	**0.0591**	**0.0343**	**0.0276**	**0.0480**	**2.7536**
	RMSE	**99.06**	**74.18**	**62.00**	**104.71**	**238.38**

CONCLUSIONS & MODEL INFERENCES (REGRESSION)

- ✓ Reading from the regression models Random Forest (RF) and KNN were superior models in predicting appliance energy use. The bias error as measured by the root mean squared error (RMSE) was considerably low averaging 62.00 and 74.18 respectively.
- ✓ The two models show good improvement from the stats model score which posted r-squared of 0.152 compared to RF (0.634) and KNN (0.4760).
- ✓ However, Cross validation score gave a negative r-square and very high variance between training sets as measured by K-fold score and K-fold standard deviation(@cv=5), without the benefit of an intercept, the these two regression models could do worse than the sample mean in terms of tracking the dependent variable.
- ✓ Linear Regression (LR) model was acceptable but the RMSE was a bit high and the explained relationship was very weak as shown by r-square of 0.1394.
- ✓ SVR and Adaboost could not predict the relationship and therefore scores obtained were not interpretable in statistical sense.

MODEL RECOMMENDATIONS & SUGGESTIONS FOR FURTHER IMPROVEMENTS

- ✓ Apply the LR, KNN and Random forest as the best models for predicting appliance energy use and further improve the models.
- ✓ For models with negative cross validation score and high variance high variance of predictions it is ideal to increase the training set size. increasing sample size by providing new data, which could translate into new cases or new features might improve the models.
- ✓ Improve models by filtering noise data set by applying Principal Component Analysis (PCA), use only features with high explained ratio.
- ✓ Building many regression models with different combination of variables. Then you can take an ensemble of all these models. This might help you arrive at a good model.

PROJECT_B_CLASSIFICATION PREDICTION MODELS

BACKGROUND SUMMARY

This sectional report summarizes supervised machine learning based on classification prediction models. The aim is to use Boston Crime data set to predict crime levels from level 1 to level 3. Set predictors are based on: Offence Code Group, District of occurrence, year, time, day and month of crime report/occurrence. The data set was collected for over 4 years. Crimes committed in USA are either classified into three unique levels, felony Level one, Level two and Level 3 as determined by the seriousness of the offence from minor misdemeanor to high level crimes. (see summary below)

Classification Dataset

Sr No	Dataset Name	Description	Rows	Columns
1	**Crime data set**	Based on Predictors (Offence Code Group, District of occurrence, year, time, day and month of crime report/occurrence we are predicting whether a person will commit level 1, two or three crime in USA, County of Boston.	Random (sample of 15000)	7
2	**Attribute Information**	-Offense Code Group, crime coding -District-place of crime occurrence/reporting -Year-year crime was committed -Months-months of crime report -Day-week day crime was reported -Time-Time crime was reported -Location-based on Longitude/Latitude UCR_Part-Crime level (1 to 3)	Random (sample of 15000)	7

MODEL RESULTS, ACCURACY & CROSS VALIDATION

CLASSIFICATION						
Crime data	**Cross Validation Method**	**KNN**	**Logistic regression**	**Random Forest**	**SVC**	**ADABOOST**
	Model Accuracy	**0.99**	**1**	**0.98**	**1**	**0.99**
	K-Fold Accuracy	**0.994**	**0.9958**	**0.9692**	**0.9957**	**0.9914**
	K-fold (STDV	**0.00089**	**0.0013**	**0.0046**	**0.0011**	**0.0036**
	Model Test Score	**0.99**	**0.9965**	**0.9822**	**0.9664**	**0.990**

CONCLUSIONS & INFERENCES

- ✓ KNN, Random forest and Adaboost proved to be superior models in predicting the crime classification.
- ✓ Model accuracy for the three models was between 98% and 99% times able to classify crime levels.
- ✓ K-fold accuracy scores, test scores and accuracy score almost the same for all the three models
- ✓ K fold score stdv. shows next to zero variance between the k-fold scores, for cv=10.
- ✓ This shows that the models were 98%-99% able to learn the crime data set and correctly classified the crime levels being committed in USA.
- ✓ Logistic Regression and Support Vector Machines (SVM) proved to be overfit models in predicting crime levels. The model accuracy score was 100% and thus not acceptable models.

MODEL RECOMMENDATIONS & SUGGESTION FOR FURTHER IMPROVEMENT (CLASSIFICATION MODELS)

- ✓ Recommend use of KNN, Random Forest and Adaboost models for the crime data set prediction and improve the models by selecting all the best hyperparameters through GridsearchCV
- ✓ **Improve model by adding more data**-for models with high variance high variance of predictions it is ideal to increase the training set size. Try increasing your sample by providing new data, which could translate into new cases or new features
- ✓ Algorithm tuning and use of multiple algorithm improves outcome of model learning process
- ✓ **Use of ensemble-**apply bagging (to reduce variance) and boosting (adjust the weights based on the last classification)
- ✓ Transforming before multilevel modelling (thus attempting to make coefficients more comparable, thus allowing more effective second-level regressions, which in turn improve partial pooling
- ✓ **Feature engineering**-apart from transformations, creating new variables out of existing variables is also very helpful. to create new features that improve the model's performance. Every new feature can make guessing the target response easier. Automatic feature creation is possible using polynomial expansion or the support vector machines class of machine learning algorithms. Support vector machines can automatically look for better features in higher-dimensional feature spaces in a way that's both computationally fast and memory optimal.

***……END*----------**