DATA MINING_PROJECT_ROBSON_MUWANI

PROJECT_A_LINEAR_ REGRESSION

BACKGROUND SUMMARY

This report provides model prediction for Melbourne House prices from 2016-2018. The buyers, financiers and the government regulators, in Melbourne City, Australia are all hoping for an orderly upturn in house pricing. Recent price gains suggest the bounce is proving to be anything but orderly - prices are threatening to spin out of control. The report seeks to explain the causes of the recent property price run in Melbourne City. Results of data analyzed in the model reveal that property price variation was 52.7% explained by distance from the city, number of rooms, number of bathrooms, land-size, car parking space, building area as indicated in the prediction analysis. Therefore there is correlation between property price and the independent/explanatory variables although the relationship is moderate. As part of the recommendation, stakeholders must further investigate other factors driving the property market in Melbourne.

OBJECTIVES, SCOPE & METHODOLOGY

1.1 Objectives:

-design model to predict house prices based on Melbourne City data set
-Seek to find if analysed variables are key drivers of property prices in Melbourne City
-Make recommendation on best predictive model and /or further investigate other drivers of property market.

The researcher conducted a predictive muilti-linear regression (MLR) model analysis for Melbourne City 2016-2018. The specific focus was to predict the key factors causing the property price run in Melbourne by analysing 10 selected variables (attributes).

DATASET DESCRIPTION (Attribute Information):

Y(Target Variable)	Property Prices
X1	# of rooms
X2	Distance from City
X3	# of bedrooms
X4	# of bathrooms
X5	Car parking space
X6	Total land size
X7	Building Area
X8	Year house built
X9	Location: latitude
X10	Location: Longitude

From the dataset I excluded land size due very weak correlation close to zero and number of bedrooms since this measure was captured in variable number of rooms and had an alpha value greater than 0.05. Number of bedrooms and rooms had a very high correlation and was eliminated in the model.

CORRELATION ANALYSIS

Correlation analysis shows that number of rooms, distance from the city, building area, bedrooms had a fairly moderate impact/positive correlation (ranging from 0.2 to 0.5) with house prices compared with the rest of other variables. Price was more sensitive to Number of rooms and distance (0.4 to 0.5 correlations). Land size and year built have a weak or insignificant impact on property prices as depicted by very weak correlation from the model

MODEL VALIDATION (MLR)

Model validation using kfold r2 score was 0.5573 compared with to 0.527 from the model summary. The test r2 score was 11% lower than the kfold validation score.

CONCLUSIONS & RECOMMENDATIONS

Results of data analyzed in the model that the property price variation was 52.7% explained by distance from the city, number of rooms, number of bathrooms, land-size, car parking space, building area as indicated in the prediction analysis. For each unit change in rooms, bathrooms, car parking space, building area had a positive impact on property pricing. Therefore there is correlation between property price and the independent/explanatory variables although it ranges from weak to moderate. The MLR was not the best model suit for property price prediction due to very high root mean squared error of $**450,409.** Since the model RMSE score is too high than what is generally acceptable in property market., I therefore recommend use of other models like Autoregressive Model,S& P/Case-Shiller Model ,GAM regression and Tree-based algorithms that minimize squared error

DATA MINING_PROJECT_ROBSON_MUWANI

PROJECT_A_LOGISTIC_ REGRESSION

PROJECT REPORT SUMMARY

Telecom companies are concerned about the number of customers leaving their services. Industry operators in the telecoms space need to understand who is leaving and why they leaving. The purpose of this project was to examine Watson dataset and design Classification model to predict customer churn rate/ customer attrition. From the results the model correctly predicted churners and non-churners 80% of the time (accuracy). Recall rate was 91% for non-churners and 50% of churners i.e. the model was more biased in predicting non churners compared to those that churned because of high frequency in those who did not churn.

OBJECTIVES, SCOPE & METHODOLOGY

The researcher conducted a predictive logistic regression model analysis for Watson Historical data set. The specific focus was to design predictive model on customer churn rate given the various services provided. The information on churn rate (Yes/No) for each customer was recorded and analysed against all the key services provided. The variable information is clearly defined as below:

Attribute Information:

gender	-Sex identity	MultipleLines	-Yes/No
SeniorCitizen	- Yes/No	InternetService	-Type of Services, DSL or Fiber Optic
Partner	-Yes/No	OnlineSecurity	-Yes/No
Dependents	-# of dependents	OnlineBackup	- service availability defined by Yes/No
Tenure	-Period with Company	DeviceProtection	- service availability defined by Yes/No
Churn(Target variable)		Yes/No	

******-all the variables were encoded to binary for reading the dataset in the model

MODEL RESULTS, ACCURACY & VALIDATION

-the model cross validation @ CV=10 precision score was 65% and predictive accuracy score from test set at 80%

-The model reported accuracy of 80%, meaning out of the 2110 observations used in the test, 80% were correctly predicted whether or not somebody churned.

[No] Row Analysis: 1411 people did not churn and were correctly predicted, while 144 who did not churn were predicted to have churned. (Recall-91%)

[Yes] Row Analysis: 278 people who did churn and were correctly predicted, while 277 who did churn were predicted to have not churned. (Recall-50%)

Precision: Non Churners (84%)-model correctly predicted did not churn 1411 versus total of 1411+277

Precision: Churners (66%)-model correctly predicted those who churned 278 versus total of 278+144

CONCLUSIONS & RECOMMENDATIONS

The model predicted a total of 1688 customers as not churn but the actual figure was 1555 while those that churned was predicted at 422 compared to 555 the actual numbers of customers who churned.

- Model accuracy of 80% is quite acceptable for predicting churn rate.
- Use model to predict customers at risk of churn
- Companies to proactively re-engage customers before they leave.

To further improve the model I recommend use of the Akaike information criterion (AIC) which is an estimator of out-of-sample prediction error and thereby relative quality of this statistical models for a given set of data. The AIC is less noisy because there is no random component in it, whereas the out-of-sample predictive accuracy is sensitive to which data points were randomly selected for the estimation and validation (out-of-sample) data.

------END--------