**DATA ANALYSIS FOR PURCHASES CHALLENGE**

1. **Methodology for exploring data was to look at summary stats, and regression analysis in excel to get a feel for what the data might show.**

|  |  |  |
| --- | --- | --- |
| Purchased | No Purchase | Total |
| 143 | 257 | 400 |
| 36% | 64% | 100% |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Female = 1 | Male = 0 | Total |
| 196 | 204 | 400 |
| 49% | 51% | 100% |

**Differences in Purchase patterns between genders**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Female | Male | total |
| No Purchase | 130 | 127 | 257 |
| Purchase | 66 | 77 | 143 |
| Total | 196 | 204 | 400 |

Conditional Probability of Purchase

Prob[Purchased/given female] = 66/196 = 34%

Prob[purchased/given male] = 77/204 = 38 %

Conditional probability for Males is a little higher than females

**Linear Probability Modeling**

Model 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -0.90411 | 0.072275 | -12.5094 | 1.73E-30 | -1.0462 | -0.76202 | -1.0462 | -0.76202 |
| Gender | -0.01619 | 0.035546 | -0.45553 | 0.648974 | -0.08607 | 0.05369 | -0.08607 | 0.05369 |
| Age | 0.026611 | 0.001715 | 15.51792 | 9.17E-43 | 0.02324 | 0.029983 | 0.02324 | 0.029983 |
| EstimatedSalary | 3.84E-06 | 5.27E-07 | 7.289992 | 1.7E-12 | 2.8E-06 | 4.88E-06 | 2.8E-06 | 4.88E-06 |

Purchased = -.9041 +.0266Age + 3.84E-06EstimatedSalary -.01619Gender

Using observation 220 from Residuals

Purchase = -.9041 + .0266(59years) + 3.84E-06 (143000) - -0619(1)

Purchase = 1.21 (i.e. 121%)

1. **LPM estimates probability of purchase (insights)**

Age explanatory variable: .0266 means for every additional year probability of purchase increases by 2.66%

Salary explanatory variable: 3.84E-06 (10000) = .0380

For every $10,000 increase in salary probability of purchase increases by 3.80%

Since the sign on the gender explanatory variable is negative this means that being a male decreases probability of purchase (This contradicts conditional probability analysis).

Being a male decreases probability of purchase by 1.62%

Since linear regression is unbounded as Estimated Salary(ES) goes up probability goes up. If ES goes up to $1milliion, probability of purchase goes up to 380%.

1. **Data preparation for a successful classification model**

Since the data was munged for the most part, the only changes made were using a dummy variable (0,1) for the Male and Female binary classification.

Also the initial linear regression modeling was done using Excel

1. **Motivation/rationale for the model choices you made.**

Using three different modeling techniques Classification and Random forest, KNN, and Logistic regression I found that the model that predicts the purchase behavior of both male and female is best predicted by the Classification and Random Forest Model closely replicated by the Logistic Regression Model.

Based on the following output:

[(0.5247026447869194, 'Age'),

(0.465271178934293, 'EstimatedSalary'),

(0.01002617627878783, 'Gender')]

Age and Salary seems to predict purchase behavior better than gender. This also closely resembles the conditional probability model explained above.

1. **Model Code**

Included in the python notebook file

1. **The results of classification and relevant performance measures**

Logistic Regression Model Results

Training Data Score: 0.6233333333333333

Testing Data Score: 0.7

Classification and Random Forest results

Score 0.93

KNN results

K: 1, Train/Test Score: 0.997/0.790

k: 3, Train/Test Score: 0.897/0.850

k: 5, Train/Test Score: 0.867/0.800

k: 7, Train/Test Score: 0.857/0.830

k: 9, Train/Test Score: 0.843/0.790

k: 11, Train/Test Score: 0.840/0.800

k: 13, Train/Test Score: 0.827/0.790

k: 15, Train/Test Score: 0.810/0.810

k: 17, Train/Test Score: 0.820/0.850

k: 19, Train/Test Score: 0.793/0.870