**RESULTS AND FINDINGS**

**Exploratory Data Analysis**

The questions responses were analyzed through descriptive statistics of frequencies and visualized using bar and pie charts.

The findings were as follows:

* Most respondents use Gvault daily (245 responses), and the least frequent users were those of less than once a month (35 responses), thus it was frequently used.
* Most respondents played roles of consumer (32%), author (29%) and reviewers (27%) within Gvault.
* More than half of the respondents affirmed that training was effective (61%).
* More than half of respondents were able to complete work without help (60%).
* 56% & 23% were satisfied and very satisfied respectively, while 21% comprised of the unsatisfied group.
* 23% noticed no difference between Gvault and GDocs, while 56% claimed the efficiency of Gvault had increased as compared to that of GDocs, 15% pointed out that the efficiency had decreased.
* 78% respondent believed that Gvault had improved features as compared to GDocs whereas 18% did not, 4% did not respond.
* 46% belong to the Pharmaceutical Development and Manufacturing functional area, 25% belong to the Research & development functional area, the IT members were 15%, facilities and operations were 3%, commercial area were 3% and the other areas had 8% .
* Most respondents (51%) had offices in Foster City, and only 2% had offices in Stockley Park.
* Many respondents (44%) were of Individual Contributor job level, and Executive Leadership level were only 1%.
* Most respondents had worked at Glead for 3-4 years (27%), followed by those who had worked for more than 7 years (26%), one to two years comprised of 15%, less than a year (13%).

**Correlation analysis**

The relationship between frequency of use, effectiveness of training and satisfaction was tested to determine if the two correlate. The result of the correlation tests are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable 1** | **Variable 2** | **P-value** | **Rho value** |
| Satisfaction | Frequency of use | 0.4022 | -0.0385 |
| Satisfaction | Effectiveness | 1.946\*10-10 | 0.2879 |
| satisfaction | Complete work without help | 0.05378 | -0.08855 |

The variables were selected for correlation analysis because they had ordinal data. The p-values show that the correlation between satisfaction and effectiveness of training was statistically significant because the p-value =1.946\*10-10 is less than 0.05. There was a significant relationship between effectiveness of training and satisfaction of users, whereas frequency of use and the extent of completing work without help were not related to satisfaction

**Tests of association**

The association between roles, ease and access to documents and satisfaction was tested to determine if the two correlate. The result of the correlation tests are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable 1** | **Variable 2** | **P-value** | **Chi square statistic** |
| Satisfaction | Role | 0.1574 | 20.391 |
| Satisfaction | Ease and access to documents | 2.2\*10-16 | 148.79 |

The variables were selected for correlation analysis because they had nominal data. The p-values show that the correlation between satisfaction and Ease and access to documents was statistically significant because the p-value =2.2\*10-16 is less than 0.05. The role of a user was in any way related to satisfaction whereas the ease of access to documents in Gvault has a high association with satisfaction.

**Multinomial logistic regression**

This model is selected when the response variable is categorical with more than one class/level. In our case the satisfaction variable is categorical with 4 classes. Thus we fit a multinomial logistic regression model.

*The* ***dependent variable*** *was satisfaction.*

*The* ***predictor variables*** *selected were; (freq\_use, role, training\_instructor\_led, training\_web\_based, training\_read, no\_training, training\_effectiveness, support\_Gnet, support\_inapplication, support\_ref\_doc, support\_SOP, support\_contacted, support\_IT, complete\_without\_help, easy\_access\_documents, easy\_access\_documents , Gvault\_efficiency, Gvault\_improved)*

The variable “Gvault\_improved” was significant in predicting the odds of a user being very satisfied. Document Stewart for a functional area and Document control group member were the roles that influence the odds of satisfaction.

Most predictor variables had an effect on the odds of a user being very satisfied vs being very unsatisfied.

**Odds ratios**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Satisfaction:** *Reference level: Very unsatisfied* | **no\_training** | **training\_effectiveness (I did not take training)** | **Support\_contacted** | **easy\_access\_documents(Yes)** |
| Satisfied | 3.697612e+05 | 3.213576e-07 | 6.516624 | 7.629050 |
| Slightly Unsatisfied | 6.124279e+05 | 2.750284e-07 | 4.315527 | 1.026245 |
| Very Satisfied | 5.762262e-06 | 4.263146e+04 | 5.437626 | 39.114817 |

**Interpretation**

The odds of a user being satisfied vs being very unsatisfied were 369761.2 times more when a user receives no training. When users receive support by Contacting the Document Control or Training Group the odds of being satisfied were 6.516624 times more than being very unsatisfied. Also when there was ease of access to documents in Gvault the odds of users being satisfied were 7.62905 times more than being very unsatisfied.

**Ordinal Logistic Regression**

The survey was conducted to investigate the user experience of the Gvault QDMS post go live. The data collected comprised of mainly categorical data with ordinal and nominal scales. The dependent variable identified was “User satisfaction with overall experience of Gvault QDMS” measured in an ordinal scale ranging from very unsatisfied to very satisfied. The objective therefore was to find out the determinants of satisfaction to predict the satisfaction level of users in future. The best model that was fitted to the data was Ordinal Logistic regression because the response variable was ordinal in nature.

The model output included the p-values (to measure significance of the model and the features in predicting the level of satisfaction), t values (test statistic), and coefficients of estimation (log odds). The odds ratios (exponentials of the log odds) were however used to come up with the model equation.

Log (odds (satisfaction)) =++

Odds Ratio (satisfaction) = exp () +exp () +…+exp ()

The effect of one predictor on the odds/likelihood of satisfaction When Gvault is used at least once a week versus at least once a month, the odds of satisfaction being very satisfied are 0.9 greater than the odds of either “satisfied” , “slightly unsatisfied” or “very unsatisfied” combined. NOTE: The reference level selected was “at least once a month”

When the training is very effective versus effective training, the odds of a user being very satisfied are 4.38 greater than the odds of being either “satisfied” , “slightly unsatisfied” or “very unsatisfied” combined. NOTE: The reference level selected was “effective”

**Summary**

Most of the predictor variables were not statistically significant in predicting the level of satisfaction with the overall experience of Gvault QDMS. Training effectiveness (p-value=0.04688), Support for In-application (p-value=0.00401), complete work without (p-value=0.000395) help, and whether Gvault improved features (p-value=0.0.00479) were the variables that significantly affected satisfaction of a user.

**Random Forest Model**

The results for the ggplot comparison visualization show that most satisfied users were found to use Gvault daily. The frequency group that had least users was ‘less than once a month’. Users who had effective training had the highest likelihood of satisfaction.

The data was fit to a random forest model that produced the top best variables that could predict satisfaction and effective training was the best with no training was the worst.

**Model training**

The model registered an accuracy of 0.99953 =99.53%

The 95% Confidence Interval: (0.9832, 0.9994)

No Information Rate: 0.5621

P-Value [Acc > NIR] : < 2.2e-16

**Model testing**

**Confusion matrix**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Reference** | | | | | |
| **Prediction** | Satisfied | Slightly Unsatisfied | | Very Satisfied | | Very Unsatisfied |
| Satisfied | 26 | | 1 | | 8 | 0 |
| Slightly Unsatisfied | 3 | | 1 | | 0 | 3 |
| Very Satisfied | 1 | | 0 | | 5 | 0 |
| Very Unsatisfied | 0 | | 0 | | 0 | 0 |

The model registered an accuracy of 0.6667 =66.67%

The 95% Confidence Interval: (0.5159, 0.796)

No Information Rate: 0.625

P-Value [Acc > NIR]: 0.3313

The model was not statistically significant in predicting satisfaction because the p-value was greater than 0.05 and the accuracy was also small.

NB: This was a case of overfitting.

**XGBoost Model**

**Model training**

The model was trained in 50 iterations which aimed at reducing mean squared error (MSE).

Number of iterations: 50

Number of features: 22

The MSE was reduced from 1.292413 to 0.593132

**Model testing**

**Confusion matrix**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Reference** | | | | | |
| **Prediction** | Satisfied | Slightly Unsatisfied | | Very Satisfied | | Very Unsatisfied |
| Satisfied | 13 | | 18 | | 0 | 0 |
| Slightly Unsatisfied | 2 | | 3 | | 1 | 0 |
| Very Satisfied | 2 | | 10 | | 1 | 1 |
| Very Unsatisfied | 0 | | 2 | | 0 | 0 |

The model registered an accuracy of 0.3208 =32.08%

The 95% Confidence Interval: (0.1992., 0.4632)

No Information Rate: 0.6226

P-Value [Acc > NIR]: 1

The model was not statistically significant in predicting satisfaction because the p-value was greater than 0.05 and the accuracy was also small.

NB: This was a case of overfitting.