Name: Devika Pillai

<u>MGIS PROJECT 2</u>

Customer satisfaction is a critical aspect of any business, and the net promoter score (NPS) is a popular metric used to measure it. In this project, I developed models to predict whether a sales representative in the software product group with a college degree will score an NPS of 9 or more. This report will outline the methodology and results of my analysis, highlighting the factors that contribute to high customer satisfaction and loyalty among sales representatives in the software product group with a college degree. The findings of this study can provide valuable insights to businesses looking to improve customer satisfaction and retention. The dataset used for this analysis consists of information about sales representatives in a company. Each row represents a different sales representative, and the columns provide various pieces of information about them. Table 1 consists of a snippet of the dataset I used.

Sales_Rep	Business	Age	Female	Years	College	Personality	Certficates	Feedback	Salary	NPS	NPSnineormore
3	Software	47	1	1	Yes	Explorer	1	3.88	52600	8	0
7	Software	25	1	1	Yes	Explorer	5	3.3	68000	6	0
11	Software	53	1	11	Yes	Explorer	2	3.93	93400	8	0
14	Software	41	0	2	Yes	Explorer	3	2.17	89000	6	0
19	Software	54	1	2	Yes	Explorer	5	2.04	90200	9	1

Tab.1: First 5 rows of dataset

First, I performed exploratory data analysis on the dataset and found the following results shown in figure 1. A sales representative who has an explorer personality has a higher chance of getting an NPS score greater than or equal to 9. The chances further reduce for the Diplomat personality, further reducing for the Sentinel personality and finally, the analyst personality has the lowest chance. Females have a lower probability of receiving an NPS score of 9 or above as compared to Males. Males also have a higher salary as compared to Females. The last graph shows the highest salary for the corresponding number of years the representative has worked in the industry. The pattern according to the graph is that as the number of years the representative has worked increases, the maximum salary decreases.

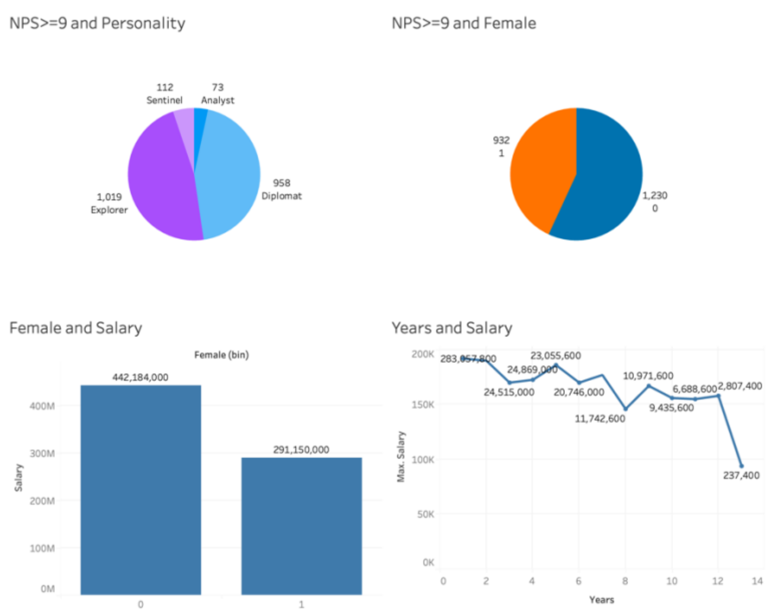

Fig.1: Dashboard consisting of graphs comparing various factors affecting NPS score

Variable	Model 1	Model 2	Model 3	Model 3 + Oversampled data
Intercept	-8.3570 (0.0000)*	-8.4356 (0.0000)*	-9.2718 (0.0000)*	-8.3666 (0.0000)*
Age	-0.0011 (0.7217)	NA	NA	NA
Female	0.2021 (0.0015)*	0.1709 (0.01089)*	0.1697 (0.0113)*	0.1545 (0.0010)*
Years	0.1892 (0.0000)*	0.3052 (0.0000)*	0.1731 (0.0000)*	0.1925 (0.0000)*
PersonalityDiplomat	1.9736 (0.0000)*	1.7892 (0.0000)*	1.7806 (0.0000)*	1.7625 (0.0000)*
PersonalityExplorer	1.9890 (0.0000)*	1.8188 (0.0000)*	1.8103 (0.0000)*	1.8777 (0.0000)*
PersonalitySentinel	0.2601 (0.1398)	0.1215 (0.5093)	0.1342 (0.4643)	0.1176 (0.2936)
Certficates	0.5392 (0.0000)*	0.5481 (0.0000)*	0.8817 (0.0000)*	0.9252 (0.0000)*
Feedback	0.6845 (0.0000)*	0.6638 (0.0000)*	1.0714 (0.0000)*	1.0703 (0.0000)*
Salary	0.0000 (0.0000)*	0.0000 (0.0000)*	0.0000 (0.0000)*	0.0000 (0.0000)*
Years:Salary	NA	-0.0000 (0.0076)*	NA	NA
Certficates:Feedback	NA	NA	-0.1158 (0.0000)*	-0.1189 (0.0000)*
Accuracy	0.813	0.81	0.817	0.771
Precision	0.36	0.356	0.372	0.791
Recall	0.653	0.642	0.665	0.494
F1 Score	0.464	0.458	0.477	0.608

Tab. 2: Logistic regression outputs

Table 2 shows the results for the logistic regression models using different predictor variables. These logistic regression models are used to predict if a customer is likely to give a high NPS score (9 or more) based on certain features such as age, gender, years of experience, personality traits, certificates, feedback, and salary. Model 1 shows that some of these features have a significant impact on the likelihood of a customer giving a high NPS score. For example, being a female, having more years of experience, having diplomat or explorer personality traits, having greater number of certificates, higher feedback scores and higher salary are associated with a higher likelihood of a customer giving a high NPS score. However, age does not seem to have a significant impact on the likelihood of a customer giving a high NPS score which is why I removed it from all other models. The model has an accuracy of 0.813, which means that it correctly predicts the NPS score for 81.3% of customers. The precision of the model is 0.36, which means that out of all the customers predicted to have a high NPS score, only 36% have a high NPS score. The recall of the model is 0.653, which means that out of all the customers who have a high NPS score, the model correctly identifies 65.3%. The F1 score is 0.464 which is less, hence it shows that the model is either biased towards precision or recall, and not both and the data is not balanced.

Model 2 show that the factors which have a significant impact on the NPS variable are being a female, years of experience, possession of certificates, having diplomat or explorer personality traits, having greater number of certificates, higher feedback scores and higher salary. On the other hand, being a Sentinel in terms of personality traits has a no significant impact on the likelihood of the representative receiving an NPS score of 9 or above. I used the interaction term of Years and Salary which shows that it is significant but has negative impact on the NPS of a representative. It shows that as the number of experience increases, the increase in NPS score is lower as the salary increases. Model 2 performed worse than model 1. The accuracy of the model is 81%, which means that the model predicted the correct outcome 81% of the time. However, the precision of the model is only 36%, which means that out of all the employees predicted as likely to recommend the company, only 36% did. The recall of the model is 64%, which means that out of all the employees who recommended the company, only 64% were correctly predicted by the model. The F1 score is low at 46%. In summary, the model has a good overall accuracy, but its ability to correctly identify those who are likely to recommend the company is not as strong, as indicated by the low precision, F1 score and moderate recall.

Model 3 shows that, being a female, years of experience, possession of certificates, having diplomat or explorer personality traits, having greater number of certificates, higher feedback scores and higher salary increase the likelihood of a customer recommending the product, while having a personality trait of Sentinel has a no significance. I used the interaction term of Certificates and Feedback which shows that it is significant but has negative impact on the NPS of a representative. It shows that as the number of certificates increases, the increase in NPS score is lower as the feedback increases. Model 3 performs the best out of the first 3 models. The accuracy of the model is 81.7%, meaning it correctly predicts the recommendation behavior of customers 81.7% of the time. The precision of the model is 37.2%, meaning that when the model predicts that a customer will recommend the product or service, it is correct 37.2% of the time. The recall of the model is 66.5%, meaning that the model correctly identifies 66.5% of the customers who would recommend the product or service. The F1 score is 0.477 for this model.

In the given dataset, the outcome variable is a binary variable with values of 0 and 1. The value 1 indicates that a sales representative has received an NPS score of 9 or higher, and 0 indicates otherwise. The number of sales representatives with an NPS score of 9 or higher is much lower than those with an NPS score below 9, resulting in class imbalance. Oversampling is a technique used to address class imbalance in binary classification problems by duplicating observations from the minority class, providing the model with more data to learn from the minority class. Being a female, years of experience, possession of certificates, having diplomat or explorer personality traits, having greater number of certificates, higher feedback scores and higher salary have a significant and positive impact and increase the likelihood of a customer recommending the product, while having a personality trait of Sentinel has a no significance. The model has an accuracy of 0.771, and it correctly predicts the outcome for 77.1% of the employees in the dataset. The precision is 0.791, the recall of the model is 0.494, which means that the model correctly identifies 49.4% of predictions for customer satisfaction. The F1 score, which is a measure of the overall performance of the model, is 0.608. This model performs the best out of all 4 models and oversampling helped improve the model's ability to predict the minority class more accurately.

Metrics	Decision Tree
Accuracy	0.793
Precision	0.309
Recall	0.573
F1 Score	0.401

Tab. 3: Decision tree output

The table 3 is the result of fitting a decision tree model to the train dataset. The tree consists of 7 nodes and has a complexity parameter of 0.039. The first node has 7889 observations, with a predicted class of 0 and expected loss of 0.218. The variable importance shows that Salary is the most important variable, followed by Certificates, Feedback, and Personality. Each node is split based on the variable and value that results in the largest improvement in the Gini index, which measures the degree of impurity of the node. The tree is pruned using cross-validation to select the optimal complexity parameter that minimizes the cross-validated error. The tree can be interpreted as follows: If Salary is less than 87700 and Certificates is less than 3.5, the predicted class is 0. If Salary is greater than or equal to 87700 and Certificates is less than 3.5, the predicted class is 0 for 75% of the observations, and 1 for 25% of the observations. If Certificates is greater than or equal to 3.5, the predicted class is 0 for 68% of the observations and 1 for 56% of the observations. The model has an accuracy of 0.793, and it correctly predicts the outcome for 79.3% of the employees in the dataset. The precision is 0.309, the recall of the model is 0.573, which means that the model correctly identifies 57.3% of predictions for customer satisfaction. The F1 score, which is a measure of the overall performance of the model, is 0.401.

Variable	Cross validation
Intercept	-9.4313 (0.0000)*
Female	0.2065 (0.0042)*
Years	0.1882 (0.0000)*
PersonalityDiplomat	1.8135 (0.0000)*
PersonalityExplorer	1.8388 (0.0000)*
PersonalitySentinel	0.1650 (0.4025)
Certficates	0.8850 (0.0000)*
Feedback	1.0339 (0.0000)*
Salary	0.0000 (0.0000)*
Certficates:Feedback	-0.1121 (0.0001)*
Accuracy	0.813
Precision	0.648
Recall	0.391
F1 Score	0.487

Tab. 4: Cross validation results

I then used cross-validation and k – fold cross validation and compared their results using metrics like accuracy, precision, recall and F1 score. The results of cross-validation showed that the model had an overall accuracy of 81.3%, meaning that it correctly classified 81.3% of the employees correctly. The precision of the model was 0.648, the recall of the model was 0.391 and the F1 score, which is a weighted average of precision and recall, was 0.487, indicating that the model has room for improvement. The coefficients of the model showed that all predictor variables except the personality of Sentinel are significant predictors of NPS score. The interaction between certifications and feedback was also significant, suggesting that the effect of certifications on NPS score depends on the feedback employees receive. Overall, the model provides some insight into the factors that influence employee loyalty and advocacy, but its performance could be improved with further refinement and validation. For k – fold cross validation, I used a k value of 10 and it showed that the model had an overall accuracy of 81.7%, meaning that it correctly classified 81.7% of the employees correctly. The precision of the model was 0.633, the recall of the model was 0.398 and the F1 score, which is a weighted average of precision and recall, was 0.488, indicating that the model has room for improvement.

Metrics	K – fold cross validation
Accuracy	0.817
Precision	0.633
Recall	0.398
F1 Score	0.488

Tab. 5: K – fold cross validation results

In conclusion, I developed various models to predict whether a sales representative in the software product group with a college degree will score an NPS of 9 or more. The models were compared using accuracy, precision, recall and F1 score. The results suggest that the employees that being a female, having more years of experience, having diplomat or explorer personality traits, having greater number of certificates, higher feedback scores and higher salary are associated with a higher likelihood of a customer giving a high NPS score as these are significant factors. It also shows that age is not a factor for predicting the NPS score and having a Sentinel personality does not matter either. It may be helpful to gather more data on customer satisfaction or NPS scores to improve the accuracy of the model. In real-world scenarios, this model could be used by a company to identify factors that are important for customer satisfaction and retention, and to predict which customers are most likely to recommend the product or service.

-X-