PREDICTIVE MODELLING PROJECT REPORT

**Executive Summary:**

This report indicates the development and selection of a predictive model for Western Alliance Bank to optimize the effectiveness of marketing strategies targeting term deposit subscriptions. Using the principles of CRISP-DM methodology described by Chapman, P., Clinton, J., Kerber, R in CRISP-DM 1.0: Step-by-step data mining guide, we focus on building a model that predicts if a customer will subscribe to the term deposit or not.

**1. Business Understanding:**

Objectives:

* To use predictive analytics for improving term deposit subscription rates.
* To ensure marketing efforts by targeting customers with a high probability to subscribe.
* To achieve a cost-effective marketing approach by focusing the efforts on reaching out to potential customers who are most likely to be interested, rather than spending resources on those who are less likely to engage with the term deposit subscription.

Problem Statement:

The bank uses historical customer data, including demographic, financial and campaign interaction details to predict term deposit subscription likelihood, thereby enhancing marketing campaign focus and effectiveness.

**2. Data Understanding:**

Data Overview:

The dataset Bank.xlsx comprises 16 predictive input variables reflecting a spectrum of customer information and historical data.

The binary output variable captures the subscription status to term deposits.

Data Quality Assessment:

* Rigorous assessment on the data indicates that the dataset's integrity and suitability for predictive modelling.
* Represents an array of customer profiles and interactions.

**3. Data Preparation:**

Data Cleaning and Transformation:

* Systematically cleaned data to ensure data quality, as guided by principles outlined in *Shmueli, Patel, and Bruce (2005) - Data Mining In Excel: Lecture Notes and Cases*.
* This was done by checking for missing values in each columns, checking the data type for each column, checking for any duplicated rows of data.
* Combined Poutcomes ‘other’ with ‘unknown’, so we have data distributed evenly for Poutomes column.
* Removed rows where 'duration' is less than 5 seconds which indicates that the customer wasn’t interested in the call (inclusion of these rows could lead to incorrect predications of the output variable).
* Checked for outliers in:

1. **Age**: Age outliers could be young adults or senior citizens. Both groups have unique banking behaviours where the young adults potentially having less financial stability and the senior citizens possibly having more savings. Hence, checked and removed ages which isn’t an appropriate age and didn’t remove the other outliers.
2. **Balance**: High balance outliers may represent high net worth individuals, who are often the target of bank marketing campaigns due to their higher financial capability to invest in products like term deposits. Hence, didn’t remove any outliers here to preserve the information of extreme values.
3. **Duration**: Call duration is an indication of customer engagement. Short calls may mean disinterest or unsuccessful contact attempts, while very long calls might represent successful engagements or complex customer inquiries. Hence not removing them to preserve this information but only removed rows with duration less than 5 seconds as mentioned above.
4. **Campaign**: A high number of contacts could either mean good marketing efforts or a system error. Frequent contacts may irritate customers, leading to lower success rates. Hence, as it’s unable to access if the repeated contacts are due to any system errors or business strategies, so to make sure there’s no loss of data, outliers aren’t removed here.
5. **Pdays**: Clients not contacted for a very long time (high 'pdays') could represent a re-engagement opportunity or outdated records. Hence, if the cases aren’t system errors and represent a part of the customer base, keeping the outliers would benefit us in not losing valuable information.
6. **Previous**: High values for previous contacts could indicate a strong past relationship or an error. Hence, if the cases aren’t system errors and represent a part of the customer base, keeping the outliers would benefit us in not losing valuable information.

* Checking what columns can be removed through data distribution:

1. **Contact** Column: The distribution of the 'contact' variable in the dataset shows that there are 2896 instances for Cellular, 1324 instances for Unknown and 301 instances for Telephone.

We can see that the distribution of the 'contact' variable is quite imbalanced. A large majority of the data is concentrated in the 'cellular' category, with 'telephone' contacts representing a small fraction. This imbalance might limit the variable's effectiveness in providing predictive power for the model. Hence, removing a variable with such an imbalanced distribution can help in reducing biases in the model. Since 'contact' is not a significant predictor, the removal of it can make the model more robust by focusing on more relevant factors. Removal of this column is also justified through business context as focusing on the communication method might not align well with the bank's goals. It could be more beneficial to focus on customer demographics, financial behaviour and other more directly relevant factors.

1. **Marital Status** Column: The distribution of the 'marital status' variable reveals the following counts: 2,797 instances of 'married', 1,196 instances of 'single', and 528 instances of 'divorced'. Hence, not removing this column as data has a balanced distribution.
2. **Education** Column: The distribution is as follows: 2,306 instances of 'secondary' education, 1,350 instances of 'tertiary' education, 678 instances of 'primary' education, and 187 instances categorized as 'unknown'. Hence, not removing this column as data has a balanced distribution.

* Used one hot encoding to encode the categorical variables to 1s and 0s for the analytical readiness to feed the data in XL Miner classification modelling.
* Dataset split:

1. **Standard partitioning** (SP) with 50% training, 30% validation, 20% testing.
2. **Partitioning with Oversampling** (OS) with a balanced training set and 25% of validation data reserved for testing.

The reason I’ve evaluated the models on the partitioning with oversampling is because the given dataset is small (after data cleaning mentioned above, we have approximately 4500 rows of data) and another significant reasoning is due to the percentage of successes in the output variable is very low in the dataset (11.8% of yes’s). So using partitioning with oversampling will train data with 50% (here) of successes (yes’s) to avoid class imbalance.

**4. Modelling:**

Model selection and deployment strategies were informed by *Kuhn and Johnson (2013) and Provost and Fawcett (2013) - "Applied Predictive Modelling.”*

Various Models Used:

A simple chart to depict the various models used:

A diagram of a company

Description automatically generated

Model Selection:

* Models with 100% accuracy may be considered for overfitting concerns and lack of generalizability. These models cannot be primary choice of deployment.
* Selected prediction on models' capacity for binary classification and adaptability to the campaign's unique dataset.

Model Deployment:

* Rigorous training conducted on standard partitioning and partitioning with oversampling to enhance model precision and accuracy.

**5. Evaluation:**

Evaluation criteria were adopted following the frameworks suggested in *Kuhn and Johnson (2013) - "Applied Predictive Modelling.”*

Performance Metrics:

* **Precision** is the best metric to ensure marketing efforts are correctly concentrated on the most promising prospects, thus maximizing return on investment. Precision is the ratio of true positives (correctly identified subscribers) to all predicted positives (all instances where the model predicted a subscriber, correctly or not). A high precision means that when the model predicts a customer is a subscriber, it is very likely to be correct.
* Precision needs to be balanced with **sensitivity**, which is the ability of the model to identify all actual subscribers.
* Other metrics considered include **accuracy, specificity** and **F1 score.**
* Evaluation is conducted on standard partitioning and partitioning with oversampling test scores.

Model Results:

Interpretation of the results were understood by *Shmueli, G., Patel, N.R. and Bruce, P.C., 2005. Data Mining In Excel: Lecture Notes and Cases.*

1. **Logistic Regression:**

a. Standard Partitioning

Performed Logistic Regression Modelling and got the test scores as below:

A screenshot of a computer

Description automatically generated  
The overall accuracy of the model is quite high, this is due to its ability to predict the 'no' class well (high specificity). The model struggles with the 'yes' class, leading to a low sensitivity (recall) and a moderate precision. This imbalance in performance is due to a class imbalance in the dataset, where there are significantly more 'no' instances than 'yes' instances. The F1 score being less than the accuracy suggests that the model is better at predicting the majority class than the minority class. Hence, performing the modelling with oversampling to check the test scores.

b. Partitioning with Oversampling

Performed Logistic Regression Modelling and got the test scores as below:

A screenshot of a test results

Description automatically generated

The model is good at identifying both 'no' and 'yes' cases, with higher sensitivity than precision, meaning it is better at catching all the positive cases but at the expense of incorrectly labelling some negative cases as positive.

The lower precision suggests that while the model can identify most of the 'yes' cases, it also misclassifies a significant number of 'no' cases as 'yes'.

The accuracy is relatively high, but the model's ability to precisely predict 'yes' cases is limited, as shown by the precision and F1 score. This could be problematic if the cost of a false positive is high.

2. **Classification Trees:**

a. Standard Partitioning

Performed Classification tree Modelling and got the test scores as below for fully grown tree:  
A screenshot of a computer

Description automatically generated

While the overall accuracy of the model is quite high (87.28%), the low sensitivity (recall) indicates the model is not very good at identifying the 'yes' cases. The high specificity suggests the model is very effective at predicting 'no' cases but at the cost of missing a significant number of 'yes' cases (high false negative rate). The F1 score is quite low (39.79%), which indicates that the model's predictive performance is not balanced—the model is biased towards predicting 'no'. The model's relatively low precision and F1 score, especially considering the high specificity, suggests that it may not be the best model for contexts where the correct identification of 'yes' cases is critical.

b. Partitioning with Oversampling

i. **Fully Grown Tree:**

A screenshot of a computer

Description automatically generated

The model has moderate accuracy, however, it has a relatively balanced sensitivity and specificity, which means it is reasonably good at detecting both 'yes' and 'no' classes. The precision is quite low, indicating a high proportion of false positives among the cases predicted as 'yes'. The F1 score, which balances precision and recall, is not particularly high, suggesting that the model's ability to predict positive cases correctly is limited. This is likely due to the large number of false positives.

ii. **Best Pruned:**

A screenshot of a computer

Description automatically generated

The model has a relatively high overall accuracy, indicating it is performing well in classifying the data. The model's specificity is high, meaning it's good at predicting the 'no' class but still has some room for improvement in minimizing false positives. The sensitivity is moderately high, which is positive as it means the model has a decent capability in catching 'yes' outcomes. The precision is almost moderate, indicating that there are a relatively high number of false positives in the predictions. The F1 score, while not exceptionally high, is reasonable and suggests that the model has a balanced performance between precision and recall.

iii. **Minimum Error Tree:**

A screenshot of a computer

Description automatically generated

The model has a high overall accuracy, indicating it performs well in classifying the data. The model's specificity is high, which is good for predicting the 'no' class, but there is still room for improvement in reducing the number of false positives. The sensitivity is moderately high, which means the model has a decent capability in catching 'yes' outcomes. The precision is moderate, indicating that there are quite a few false positives in the predictions. The F1 score is moderately good and suggests that the model has a balanced performance between precision and recall.

3. **K-Nearest Neighbours (KNN):**

a. Standard Partitioning

A screenshot of a computer

Description automatically generated

This model shows a high overall accuracy (92.36%) but seems to have a problem with false negatives (high error rate for the 'yes' class, indicating poor recall). The specificity is excellent, meaning it is very good at identifying the negative class, but the sensitivity or recall for the positive class is quite low (around 41.51%), indicating that the model is missing a significant number of actual positive instances. Precision is relatively high, so when the model predicts 'yes', it is correct about 86.28% of the time. However, the F1 score is moderate at 0.56051, reflecting the imbalance between precision and recall.

b. Partitioning with Oversampling

A screenshot of a test results

Description automatically generated

The model has a decent overall accuracy, but it's not very high, suggesting room for improvement. Specificity is good, meaning the model can identify 'no' cases reliably. Sensitivity is also good, which indicates that the model has a good ability to identify 'yes' cases. However, the precision is quite low, suggesting that the model incorrectly predicts 'yes' quite often when it's actually 'no'. The F1 score is moderate, indicating that the balance between precision and recall is not optimal.

4. **Neural Network:**

a. Standard Partitioning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing % Error** | **Testing % Sensitivity** | **Testing % Specificity** | **Testing % Precision** | **Testing % F1-Score** |
| 11.72566 | 0 | 100 | N/A | N/A |

your neural network model seems to have a significant bias towards predicting negative outcomes, doing so perfectly (specificity of 100) but failing to correctly identify any positive cases (sensitivity of 0). Since, there’s a N/A for precision and F1 score, we can conclude that the Neural Network model is not applicable for our case.

b. Partitioning with Oversampling

Same results as the above.

5. **Ensemble Bagging:**

a. **Logistic Regression Ensemble Bagging:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model has fairly high accuracy and specificity, its sensitivity is low, which means it struggles with correctly identifying true positive cases. This imbalance suggests that the model may be conservative in predicting the 'yes' class. The relatively low F1 score also supports this, as it indicates a trade-off between precision and recall is not optimal.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The model has a decent overall accuracy and a balanced sensitivity and specificity, its precision is low. This implies that when the model predicts a 'yes', it is correct only about one-third of the time. Therefore, while the model is relatively good at identifying positive cases among all actual positives, many of the cases it predicts as positive but are actually negative.

b. **KNN Ensemble Bagging:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model shows a high specificity but a low sensitivity, indicating that it is better at predicting negative instances than positive ones. The high error rate for the positive class and the low F1 score suggest that the model is not performing well in terms of identifying the positive cases accurately.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The model shows excellent performance on the test data with high values across all metrics.

c. **Decision Tree Ensemble Bagging:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model has a high specificity but a low sensitivity, which indicates that it is better at predicting negative instances than positive ones. The high error rate for the 'yes' class and the moderate F1 score suggest that the model is not very effective at correctly identifying positive cases.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The low precision and moderate F1 score suggest that while the model is good at getting positive cases (high recall), it also makes a significant number of incorrect positive predictions in comparison to the actual number of positive cases. This indicates that this model is biased towards predicting the 'yes' class.

6. **Ensemble Boosting:**

a. **Logistic Regression Ensemble Boosting:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model has high specificity but low sensitivity, indicating it is much better at predicting negative cases than positive ones. The relatively low F1 score and moderate precision suggest that the model does not balance well between recall and precision.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The model's accuracy is moderate and it has similar specificity and sensitivity, meaning it is equally good at identifying both negative and positive cases. However, the precision is quite low, which means a large proportion of the cases predicted as positive are actually negative. The moderate F1 score further reflects the imbalance between the model's recall and precision.

b. **KNN Ensemble Boosting:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model demonstrates high specificity but very low sensitivity, indicating it is much better at predicting negative instances than positive ones. The low sensitivity means the model misses a significant number of actual positive cases. The precision is also relatively low, which means a considerable portion of the positive predictions made by the model are false. Consequently, the F1 score is low, suggesting that the model's ability to balance precision and recall is not optimal.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The model has a moderate sensitivity, indicating a fair ability to identify positive instances. However, the precision is quite low, which means that many positive predictions are incorrect. The F1 score, which balances precision and recall, is also moderate, reflecting the trade-off between these metrics. The accuracy is just above 71%, which is not very high for a classification model.

c. **Decision Tree Ensemble Boosting:**

Standard Partition:

A screenshot of a computer

Description automatically generated

The model performs well in terms of specificity, meaning it's quite good at predicting the negative class. However, its performance on the positive class is not as strong. The sensitivity is less than 50%, indicating that more than half of the actual positive cases are being missed. Similarly, the precision is under 50%, which means that a significant portion of the positive predictions are false positives.

Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

The model has a relatively high sensitivity, which means it's quite effective at identifying positive cases. However, the precision is low, which indicates that many of the cases the model predicts as positive are actually negative. This is also reflected in the high false positive rate in the confusion matrix. The F1 score, which combines precision and recall into a single metric - is moderate, indicating that the model is not as effective at correctly classifying cases as it could be.

7. **Ensemble Random Tree**

a. Standard Partition:

A screenshot of a computer

Description automatically generated

This model indicates perfect performance which usually indicates overfitting.

b. Partitioning with Oversampling

A screenshot of a computer

Description automatically generated

This model is of high-performance, with nearly perfect accuracy, specificity and sensitivity. The precision is also very high, which means there are very few false positives compared to the number of true positives.

Model Comparison:

To evaluate the best model for the marketing efforts based on the specific performance metrics, we need to prioritize precision while also considering sensitivity, accuracy, specificity and F1 score.

Let’s reiterate the objectives and how our models can be based on them. Our Objectives are:

1. **Improving Term Deposit Subscription Rates**: This requires a model that not only accurately predicts potential subscribers (high precision) but also doesn't miss out on a significant number of actual subscribers (reasonable sensitivity).
2. **Cost-Effective Marketing Approach**: Precision is crucial here, as targeting customers who are most likely to subscribe will optimize resource allocation. However, it's also important to consider the overall accuracy and F1 score for a balanced approach.
3. **Improving Marketing Campaign Focus and Effectiveness**: This involves identifying the right customers (precision) without overwhelming those unlikely to subscribe (specificity) and making sure there’s a broad enough reach to capture most potential subscribers (sensitivity).

Based on the business needs, evaluating the best practical models’ results:

**KNN Ensemble Bagging (Partitioning with Oversampling) and Ensemble Random Tree:**  
A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Considering the business needs, K-Nearest Neighbours (KNN) ensemble bagging in the oversampling partition seems to be the most suitable model (left figure above). It offers high precision making sure that the marketing efforts are focused on the most promising prospects and its sensitivity is really high, which means it can also identify a good number of actual subscribers. This is crucial for a cost-effective and efficient marketing strategy.

However, it is also crucial to consider the potential for overfitting with the Ensemble Decision Trees (right figure above). If further validation confirms that the Ensemble Decision Trees model is robust and generalizes well, it might be worth considering due to its high precision and sensitivity.

In summary, for a balanced approach that meets the business objectives, the KNN ensemble bagging partitioned by oversampling seems most appropriate with a secondary consideration for the Ensemble Decision Trees model if its generalizability is confirmed.

**6. Deployment:**

The implementation strategy and performance monitoring approach are aligned with the best practices described in *Provost and Fawcett (2013) – “Applied Predictive Modelling.”*

Implementation Strategy:

* KNN Model Ensemble Bagging is recommended for deployment with performance monitoring mechanisms in place.

Performance Monitoring and Evaluation:

* Establish ongoing monitoring systems to track the performance of the KNN model in real-time. This involves setting up key performance indicators (KPIs) like precision, sensitivity, and overall accuracy.
* Regularly evaluate the model against these KPIs and adjust thresholds or parameters as needed to maintain or improve performance.

Feedback Loop Integration:

* Implement a feedback mechanism to continuously collect data on the model's predictions and actual outcomes. This data will be crucial for further refining the model.
* Use customer responses and engagement metrics to assess the effectiveness of marketing strategies driven by the model's predictions.

Model Updating and Re-training:

* Plan for periodic re-training of the model with new data to ensure that it adapts to changing customer behaviours and market trends.
* Automate the re-training process to ensure the model remains up-to-date without manual intervention.

Parallel Testing:

* Run the Ensemble Decision Trees model in parallel with the primary KNN model. This allows for a side-by-side comparison of their performances in a real-world setting.
* Use this parallel testing phase to validate the practical accuracy and generalizability of the Ensemble Decision Trees model, especially considering its 100% accuracy in testing, which might indicate overfitting.

**7. Conclusion and Recommendations**

Key Findings:

* Precision is very important in direct marketing to ensure cost-efficiency.
* KNN Ensemble Bagging Model offers the best balance between high precision and accuracy, essential for effective targeting.

Recommendations:

* Deploy KNN Ensemble Bagging Model for targeted marketing initiatives and as mentioned in the deployment section, perform parallel testing with the Ensemble Decision Trees.
* Establish a continuous improvement cycle for the model to adapt to new data and market changes as mentioned in the Deployment section.

**References**

* Winston, W.L. (2021) 'Operations Research: Applications and Algorithms', 5th edn. Cengage Learning, pp. 850-860.- **Used this reference to understand the excel solver reports.**
* Chapman, P., Clinton, J., Kerber, R., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. – **Used this reference to understand the CRISP-DM concepts and helped crafting this report.**
* Shmueli, G., Patel, N.R. and Bruce, P.C., 2005. Data Mining In Excel: Lecture Notes and Cases. Arlington, VA: Resampling Stats, Inc. – **Used this reference to understand how the results has to be evaluated for each model and also understood how each modelling is done in XL Miner.**
* Kuhn, M., Johnson, K. (2013). Applied Predictive Modelling. – **Used this to understand how can I interpret my results as this reference had practical applications.**
* Provost, F., Fawcett, T. (2013). Data Science for Business. – **Used this reference to understand the strategic considerations of business applications. It made me understand a clear perspective on how data science can be leveraged to drive business decisions.**