**Predicting the Effectiveness of a Marketing Campaign**

**Project Description**

1. **Project Title and Team Members**

Project Title is Predicting the Effectiveness of Marketing Campaign

Team Members are,

Jaynica Nunna ([Jaynicanunna@my.unt.edu](mailto:Jaynicanunna@my.unt.edu))

Rahul Siddartha ([rahulsiddarthagotti@my.unt.edu](mailto:rahulsiddarthagotti@my.unt.edu))

Sai Gayathri Makineni([SaiGayathriMakineni@my.unt.edu](mailto:SaiGayathriMakineni@my.unt.edu))

Naga Lakshmi Thota ( [NagaLakhmiThota@my.unt.edu](mailto:NagaLakhmiThota@my.unt.edu))

1. **Goals and Objectives**

**Motivation:**

The project aims to enhance marketing strategies in sectors like Real Estate, Insurance, Banking, and Finance by predicting customer responses using machine learning. This initiative seeks to reduce ineffective marketing pitches and minimise spam, thereby saving valuable man-hours and improving customer experience. Now in this tech-enabled era, many people are aware of the techniques but don't implement them due to accessibility and effectiveness. We want to bridge the gap and provide and online-learning algorithm which can train over time with the inputs from people in call-centres and be a enabler for businesses

**Significance:**

Predicting the effectiveness of marketing campaigns is a vital aspect of modern business strategy. It not only enhances decision-making but also optimises budget allocation, ensuring that resources are invested in campaigns likely to yield the highest returns. As the gap between investment and revenue widens, particularly in digital marketing, the pressure increases for marketers to deliver improved post-install campaign performance and user engagement​​. Machine learning models, like Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN), have been shown to significantly outperform traditional parametric models, with SVM identified as being particularly effective due to its ease of training, intuitiveness, and computational efficiency​​. The ability to process large datasets, identify complex patterns, and predict outcomes with high accuracy makes machine learning an invaluable tool for campaign optimization​​. Using these models, marketers can predict which campaign aspects are most influential, helping to focus efforts and resources more effectively​​.

The choice and tuning of algorithms are critical, as they can drastically affect the predictive power of the models. While some models might excel in certain scenarios, others might perform worse, highlighting the importance of selecting the right model based on the specific characteristics of the dataset at hand​​.

**Objective:**

Our objective for this project is to train different machine learning models and see how they perform with different hyperparameters and how accurately they predict the marketing campaign effectiveness. We want to prove that we can make the online version of this machine learning approach possible with a potential high accuracy. We could even use audio data from actual calls and convert them to text in the future to gather even more data and build better and accurate models.

**Features:**

* Very fast training speed (0-10) minutes
* Visualising attributes and their importance to the final outcome.

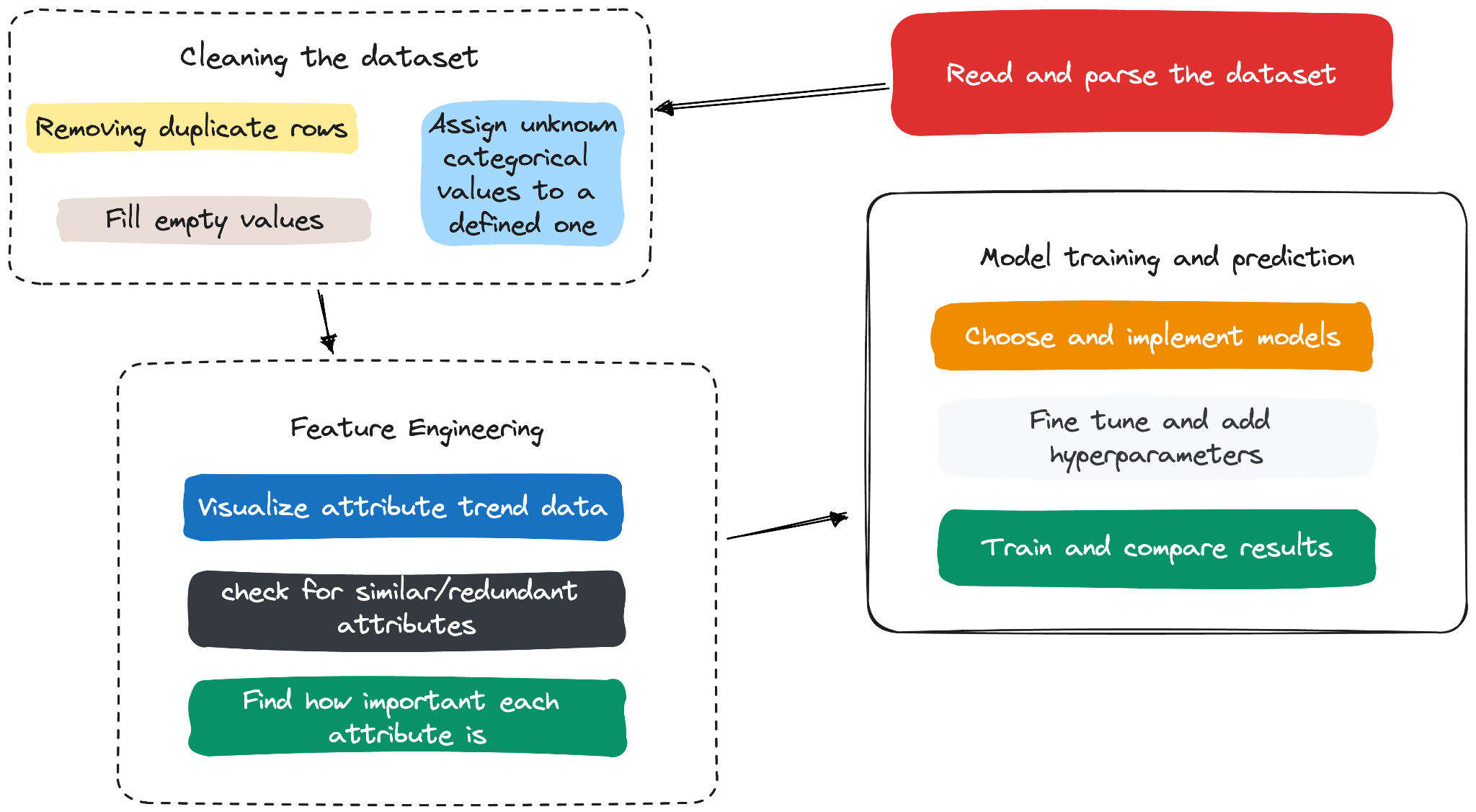
**Datasets Description**

* **Bank Marketing data from UC Irvine ML Repo:**

The Bank Marketing dataset from the UC Irvine Machine Learning Repository is a rich and comprehensive dataset that has been widely used for research in predictive analytics, particularly for understanding customer response to bank marketing campaigns. [[1]](https://archive.ics.uci.edu/dataset/222/bank+marketing) This dataset provides a combination of categorical and continuous variables, including age, job, marital status, education, default records, housing loans, personal loans, and other attributes relevant to a marketing campaign.

It's derived from direct marketing campaigns of a Portuguese banking institution, with the classification goal to predict whether a client will subscribe to a term deposit. It contains over 45,000 instances, making it suitable for complex machine learning tasks that require a substantial amount of data for training. The dataset has been pivotal in benchmarking the performance of various classification algorithms, including decision trees, ensemble methods, and logistic regression. The outcome variable is binary, simplifying the analysis to a binary classification problem, which is ideal for methods like logistic regression, but the dataset also allows for more complex analyses, such as customer segmentation and personalization strategies. The Bank Marketing dataset is particularly noted for its imbalanced class distribution, which poses a challenge and makes it an excellent resource for developing and testing strategies for imbalance learning.

**Approaches towards making predictions:**



The Bank Marketing Dataset from the UC Irvine Machine Learning Repository serves as a pivotal benchmark in the realm of predictive analytics.Processing this dataset involves a multi-step approach, as depicted in the workflow image. Initially, the dataset is cleansed by removing duplicates and imputing missing values, ensuring data quality and consistency. This is followed by reassigning ambiguous categorical values to predefined classifications, enhancing the dataset's clarity and usability.

Feature engineering is the next critical phase, where attributes are visualised to discern trends, and redundant features are identified and addressed. This step also involves determining the importance of each attribute, a process that directly influences the model's predictive accuracy.

Subsequently, the refined dataset is used for model training and prediction. Various machine learning models are chosen and implemented, followed by fine-tuning using hyperparameters optimization to enhance model performance. The final and decisive step involves training the models and comparing their results to select the one that best predicts customer response to bank marketing campaigns. This comprehensive approach to data preparation, feature engineering, and model selection is instrumental in leveraging the full potential of the Bank Marketing Dataset for insightful predictive analysis.

**Implementation:**

In the realm of data science, the process of transforming raw data into actionable insights is a multifaceted endeavour, especially when it comes to evaluating the effectiveness of marketing campaigns. The workflow depicted in the provided image encapsulates a high-level methodology that is often employed when handling the Bank Marketing Dataset from the UC Irvine Machine Learning Repository, a popular dataset for benchmarking machine learning algorithms aimed at predictive analytics in marketing.

**Cleaning the Dataset:**

Data cleaning is the initial and a critical step that sets the foundation for accurate analysis. It involves removing duplicate entries which could skew results and filling missing values that, if left unaddressed, could lead to incorrect predictions. Categorical values that are unknown are reassigned, which helps in maintaining the consistency and integrity of the dataset for more reliable outcomes.

**Feature Engineering:**

Feature engineering is where domain expertise intersects with data science. Here, data scientists visualize attribute trends to identify which features could have the most significant impact on the model's predictions. They also check for similar or redundant attributes to avoid the curse of dimensionality, where too many features can cause the model to perform poorly[[6]](https://blog.hubspot.com/agency/predictive-analytics-buy). Determining the importance of each attribute allows for dimensionality reduction and helps focus the model on the most predictive features.

**Model Training and Prediction:**

With a polished dataset, the focus shifts to selecting and implementing machine learning models. These models are algorithms that, given input data, can make predictions or decisions without being explicitly programmed to perform the task.

* **Machine Learning Models**

The effectiveness of a marketing campaign can be predicted using various machine learning models:

**Logistic Regression:** A statistical model that, in the context of a binary outcome, estimates the probability of occurrence[[2]](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/) of an event by fitting data to a logistic curve. It is simple, interpretable, and serves as a baseline for other models.

**Regularised Greedy Forest (RGF):** An ensemble learning method that builds on decision trees. [[4]](https://saturncloud.io/glossary/regularized-greedy-forest/)RGF integrates regularisation, which helps in avoiding overfitting—a common problem where a model learns the training data too well but performs poorly on unseen data.

**Gradient Boosted Trees (XGBoost)** [**[5]**](https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=11772589)**:** An efficient and scalable implementation of gradient boosting framework. It stands out for its performance and speed and is particularly useful for large datasets.

**One-Class SVM:** It's a variant of SVM used for anomaly detection. Unlike the traditional SVM that separates two classes, the one-class SVM learns a boundary[[7]](https://www.xlstat.com/en/solutions/features/1-class-support-vector-machine) that encompasses all the data points of the single class it was trained on.

**Isolation Forest:** Another anomaly detection algorithm that isolates outliers instead of profiling and modelling the "normal" points [[8]](https://medium.com/@corymaklin/isolation-forest-799fceacdda4). It's effective in high-dimensional datasets.

**Support Vector Machines (SVM):** A robust and versatile class of supervised algorithms for both classification and regression tasks, which works well on medium and large-sized datasets[[9]](https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1183475870).

**Linear SVM:** Offers a linear decision boundary using the kernel trick.

**Quadratic and Cubic SVM:** Uses polynomial kernels of degree 2 and 3, allowing the model to fit non-linear relationships.

Each model has its strengths and is chosen based on the specific characteristics of the data and the analytical goals.

* **Visualisation Libraries:**

Alongside these models, visualisation libraries play a significant role:

**Matplotlib:** A plotting library that offers comprehensive graphs and figures, a valuable tool for visualising trends and relationships in data.

**Seaborn:** Based on Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.

These libraries help data scientists to visualise and understand data and the underlying trends, which is crucial before any predictive modelling can take place.

* **Hyperparameter Tuning and Results Comparison**

Choosing the right hyperparameters for these models can significantly affect their performance[[10]](https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/). Techniques like cross-validation and grid search are often used to find the optimal settings. The performance of these models is then compared using evaluation metrics like ROC-AUC scores [[3]](https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc), which provide insight into the true positive rate versus the false positive rate, crucial for classification tasks, especially with imbalanced datasets.

In **conclusion**, employing machine learning models to predict the effectiveness of marketing campaigns involves an iterative process of cleaning the data, engineering features, choosing the right models, and tuning them to perfection. Each model has its particular use case, and the choice of model depends on the dataset's characteristics and the problem at hand. Visualisation tools aid in this process by allowing for an intuitive understanding of the data, which in turn informs better model selection and tuning, leading to more reliable predictions and actionable insights.

**Implementation Status Report :**

We have implemented multiple machine learning models and have learnt a lot about the problem statement. We have tried other approaches but with minimal success. We do not believe deep learning might be helpful unless we have image or audio data as part of our initial input dataset.

**Work completed**

We have trained the models using most of the promising algorithms and found the below results [[3]](https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc)

| ML Model | ROC-AUC Score |
| --- | --- |
| Baseline | 0.87 |
| Reguralized Greedy Forest (RGF) | 0.92 |
| Gradient Boosted Trees with XGBoost | 0.922 |
| One Class SVM | 0.63 |
| Isolation Forest | 0.58 |
| Linear SVM | 0.85 |
| Quadratic SVM | 0.79 |
| Cubic SVM | 0.86 |

We find that RGF and Gradient Boosted Trees performed better than the other models and might even outperform with better quality of data.

**Work to be completed**

We still need to implement the online version of the machine learning model which will incrementally train the models based on user inputs in real-time and get better at predicting everyday. This is a complex task to implement but for now we have used and implemented significant models to give this a try in the future.

**Issues:**

The data is not comprehensive and more data would be really helpful to the model. We might even collect non-transactional data like call recording and economy, bank details and credit score which helps us to make even better predictions. Few of the attributes were found to be redundant, reducing the information the model can access.

**Responsibility:**

* **Jaynica Nunna** is responsible for Data Preprocessing like removing duplicates and imputing missing values etc, some part of Documentation and Model Implementation such as One-class SVM.
* **Rahul Siddartha** is responsible for Research, Data Analysis, Feature Engineering, Model Tuning , Model implementation such as Logistic Regression, RGB Classifier.
* **Sai Gayathri Makineni** is responsible for Experimentation, Data Preprocessing line assigning unknown categorical values to a defined one etc, Analysis of Metrics and Result Interpretation
* **Naga Lakshmi Thota** is responsible for Model Implementation such as Isolation model, SVM models such as linear SVM model, Quadratic SVM model and cubic SVM model, Qualitative Analysis, some part of documentation, Testing and Validation

**GITHUB Link:**

<https://github.com/Nagalakshmi344/ML-PROJECT.git>

**References**

1. UCI Machine Learning Repository. (n.d.-h).

<https://archive.ics.uci.edu/dataset/222/bank+marketing>

2. “Logistic Regression: Equation, Assumptions, Types, and Best Practices.” *Spiceworks*, <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

Accessed 19 Nov. 2023.

3. “Classification: ROC Curve and AUC | Machine Learning.” Google for

Developers,<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc> Accessed 13 Oct. 2023.

4. *Regularized Greedy Forest | Saturn Cloud*. 14 Apr. 2023, <https://saturncloud.io/glossary/regularized-greedy-forest/>.

5. “Gradient Boosting.” Wikipedia, 26 Sept. 2023. Wikipedia,

<https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=11772589>

6. How to Use Predictive Purchase Behavior Modeling to Understand Consumers.

<https://blog.hubspot.com/agency/predictive-analytics-buy>

7. “One-Class Support Vector Machine.” *XLSTAT, Your Data Analysis Solution*, <https://www.xlstat.com/en/solutions/features/1-class-support-vector-machine>.

Accessed 19 Nov. 2023.

8. Maklin, Cory. “Isolation Forest.” *Medium*, 15 July 2022, <https://medium.com/@corymaklin/isolation-forest-799fceacdda4>.

9. “Support Vector Machine.” *Wikipedia*, 4 Nov. 2023. *Wikipedia*, <https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1183475870>.

10. Pandian, Shanthababu. “A Comprehensive Guide on Hyperparameter Tuning and Its Techniques.” Analytics Vidhya, 22 Feb. 2022, <https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/>.