**Capstone Project - Customer Churn**

**Table of Contents**

**1. Problem Statement**

**2. Project Objective**

**3. Data Description**

**4. Data Pre-processing Steps and Inspiration**

**5. Choosing the Algorithm for the Project**

**6. Motivation and Reasons for Choosing the Algorithm**

**7. Assumptions**

**8. Model Evaluation and Techniques**

**9. Inferences from the Same**

**10. Future Possibilities of the Project**

**11. Conclusion**

**12. References**

**Problem Statement:**

You are the Data Scientist at a telecom company “Neo” whose customers are churning out to its competitors. You have to analyse the data of your company and find insights and stop your customers from churning out to other telecom companies

**Project Objective:**

* The objective of the project is to reduce customer churn for the telecom company "Neo" by leveraging data science and analytics. The primary goals include :
* Analyse the customer churn dataset to identify patterns and factors that contribute to customer churn in the telecom industry
* Develop machine learning models to predict which customers are more likely to churn in the future. This involves selecting appropriate algorithms, training the models, and evaluating their performance.

**Data Description :**

* Customer ID: Unique identifier for each customer.
* gender: Gender of the customer (e.g., Female, Male).
* Senior Citizen: Whether the customer is a senior citizen (0 for No, 1 for Yes).
* Partner: Whether the customer has a partner (e.g., Yes, No).
* Dependents: Whether the customer has dependents (e.g., Yes, No).
* Tenure : Number of months the customer has been with the company.
* Phone Service: Whether the customer has a phone service (e.g., Yes, No).
* Multiple Lines: Whether the customer has multiple phone lines (e.g., Yes, No phone service, No).
* Internet Service: Type of internet service subscribed (e.g., DSL, Fiber optic).
* Online Security: Whether the customer has online security (e.g., Yes, No).
* The target variable is :
* Churn: Indicates whether a customer has churned (left the service) or not (e.g., Yes, No)

**Data Pre-processing Steps and Inspiration:**

* Data was split into training and testing through train test split
* Here dependant variable was the churn column.

**DATA INSIGHTS :**

* Most of the people opted for Fiber optic internet service due to fast speed
* More outliers were found when contract was monthly basis and 2 year contract.
* Most of the contract was on monthly basis .
* Customers tend to leave the service between 10 to 65 months.
* Customers with two year contracts tend to last longer with the service.

**Choosing the Algorithm for the Project :**

* Data was trained and test on various alogirthms .
* Such as Linear Regression,Logistic Regression,Multiple Logistic Regression,DecisionTrees and Randomn Forest.
* To model was evaluated based on confusion metrics and accuracy score.
* From which multiple logistic Regression was found to be the best model.

**Motivation and Reasons for Choosing the Algorithm:**

* Logistic regression is well-suited for binary classification problems, where the goal is to predict whether an instance belongs to one of two classes. In this the outcome is binary – whether a customer will churn or no .
* Logistic regression is computationally efficient and can handle large datasets with relatively low computational resources.
* This is important in scenarios where there is a large customer base and a substantial amount of data to analyze .
* So logistic regression model was chosen

**Assumptions :**

* Observations are assumed to be independent of each other .
* The independent variables should not be highly correlated with each other. Multicollinearity can lead to unstable coefficient estimates.
* The relationship between the log-odds of the dependent variable and the independent variables is assumed to be linear .

**Model Evaluation and Techniques :**

* Confusion matrix table are used in each model to summarize the model
* While the model is evaluated based on accuracy ,more the accuracy better the model
* From that accuracy of multiple regression model was found to be the highest among all the other models.

**Inferences from the Same:**

* The model achieves an accuracy of approximately 77.46%. This suggests that, overall, the model is correct in predicting customer churn or no churn about 77.46% of the time.
* The model has a relatively high number of true positives (156), indicating that it is effective at correctly identifying customers who are likely to churn.
* There is a notable number of false positives (212), suggesting instances where the model predicts churn incorrectly.
* The model can be valuable in identifying customers at risk of churn, allowing for targeted retention efforts.

**Future Possibilities of the Project :**

* Enhance the predictive models ,this could involve adjusting hyperparameters, incorporating new features, or exploring more advanced machine learning algorithms to improve prediction accuracy.
* As new data becomes available, continuously update the model to ensure it remains relevant and effective in predicting customer churn.
* Incorporate customer feedback data into the analysis. By considering customer feedback, the model can gain additional insights into customer satisfaction and identify potential areas for improvement in products or services.

**Conclusion:**

* The customer churn prediction project used various models, including Linear Regression, Logistic Regression, Decision Trees, and Random Forest.
* Evaluation based on confusion metrics and accuracy scores revealed that Multiple Logistic Regression outperformed the others.
* The model's strength lies in its ability to capture complex relationships, making it a valuable tool for identifying customers at risk of churn.

**References :**

Customer Churn Data -Kaggle