CAPSTONE PROJECT

Customer Churn Predictions

Ryan Larsen
Data Scientist

AGENDA

1. PROJECT CONTEXT
2. DEFINE (Business & Data Science Aspects)
3. DESIGN (EDA - Exploratory Data Analysis)
4. DELIVERY (Feature Engineering & Machine Models)
5. SUMMARY, CONCLUSIONS AND NEXT STEPS
6. QUESTIONS
7. APPENDIX

BIO

EDUCATIONAL BACKGROUND

- Data Science and Artificial Intelligence Professional Certificate (IOD)
- Data Analytics Professional Certificate (Google)
- BSc: Microbiology (Massey University)

PROFESSIONAL EXPERIENCE

- Technical/Analytical Science, Manufacturing Industries, Compliance Auditing, and Quality Assurance Roles (9 yrs)

PROFESSIONAL EXPERIENCE

- Specialising in data science with technical expertise in Python, Machine Learning, and visualisation
- Equipped to analyse and predict customer churn, aligning perfectly with the customer churn analysis project

DATA SCIENCE SKILLS: Python, Machine Learning, SQL, Visualisation

1 PROJECT CONTEXT

PROJECT CONTEXT

SPIDERCOM
CONNECTING MONSTERS

- **Industry / Domain:** SpiderCom No.1 Telecommunications Company
- **Challenges Area:** Customer retention
- **Problem Area:** Identify customer behavior & Churn prediction
- **Why this area is interesting:** Drive revenue & improve customer experience
- **Previous work in the area:**

Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer churn prediction in telecommunication a decade review and classification. *International Journal of Computer Science Issues (IJCSI)*, *10*(5), 271.

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer perceptron neural networks: Modeling and analysis. *Life Science Journal*, *11*(3), 75-81.

Disclaimer: SpiderCom is a mock telecommunications company used solely for the purpose of this data science capstone project.

2

DEFINE

Business & Data Science Aspects

DEFINE
BUSINESS ASPECTS

SPIDERCOM
CONNECTING MONSTERS

- **Profile:** Leading mobile and internet service provider.
- **Market Presence:** Extensive coverage in multiple regions.
- **Stakeholders:** Management, Marketing, CX (Customer Service), Sales & Retail
- **Mission:** Enhancing customer loyalty by reducing customer churn.
- **Data Contribution:** Provided a valuable customer dataset.
- **Strategic Focus:** Conduct analysis to identify segment groups.

Disclaimer: SpiderCom is a mock telecommunications company used solely for the purpose of this data science capstone project.

"Reducing customer churn by 5% could increase profits by 25% - 95% (2)."

65% of business comes from **existing** customers (1).

“

It costs 5x as much to attract a new customer than to keep an existing one (1).

”

BUSINESS QUESTION

HOW TO REDUCE
CUSTOMER CHURN
AND ENHANCE
LOYALTY AT SPIDERCOM ?

SPIDERCOM
CONNECTING MONSTERS

DEFINE
DATA OVERVIEW

SPIDERCOM
CONNECTING MONSTERS

Type: Structured, with numerical and categorical features.

Source: Collected from SpiderCom's customer records.

Size: **7045** customers and **21** columns.

Time Period: The data represents customer behavior and attributes during the years 2021-2022.

Relevance: Essential for analysing SpiderCom customer behavior, retention, and churn patterns. This data will be instrumental in making predictions and formulating strategies to enhance customer satisfaction and loyalty.

DEFINE
DATA COLLECTION & PROCESSING

Method of Collection:
- Quality Data Gathered from customer interactions, online portals, Company Records

Data Preprocessing and Cleaning:
- Handling missing values, duplicates, outlier detection, column renaming & label encoding,

Data Integration:
- Merged billing, support, online data; joined on 'customer_id'.

Challenges and Limitations:
- Inconsistencies in 'internet_service'.
- Lack of detailed demographics.
- Potential biases in self-reported attributes.
- Unbalanced dataset, especially in the 'churn' column, requiring specific handling techniques.

DATA SCIENCE TOOLKIT

Data Preprocessing:
- Pandas (Data Manipulation)
- NumPy (Numerical Operations)
- Matplotlib & Seaborn

Modeling:
- Scikit-learn (Machine Learning)
- Bagging, Boosting, Stacking (Ensemble Techniques)
- SMOTE (minority classes)

Serialization:
- Joblib (Object Serialization)
- Pickle (Object Serialization)

Visualisation:
- Matplotlib (Plotting)
- Seaborn (Data Visualization)
-

Deployment:
- Flask (Web API Development)

DEFINE

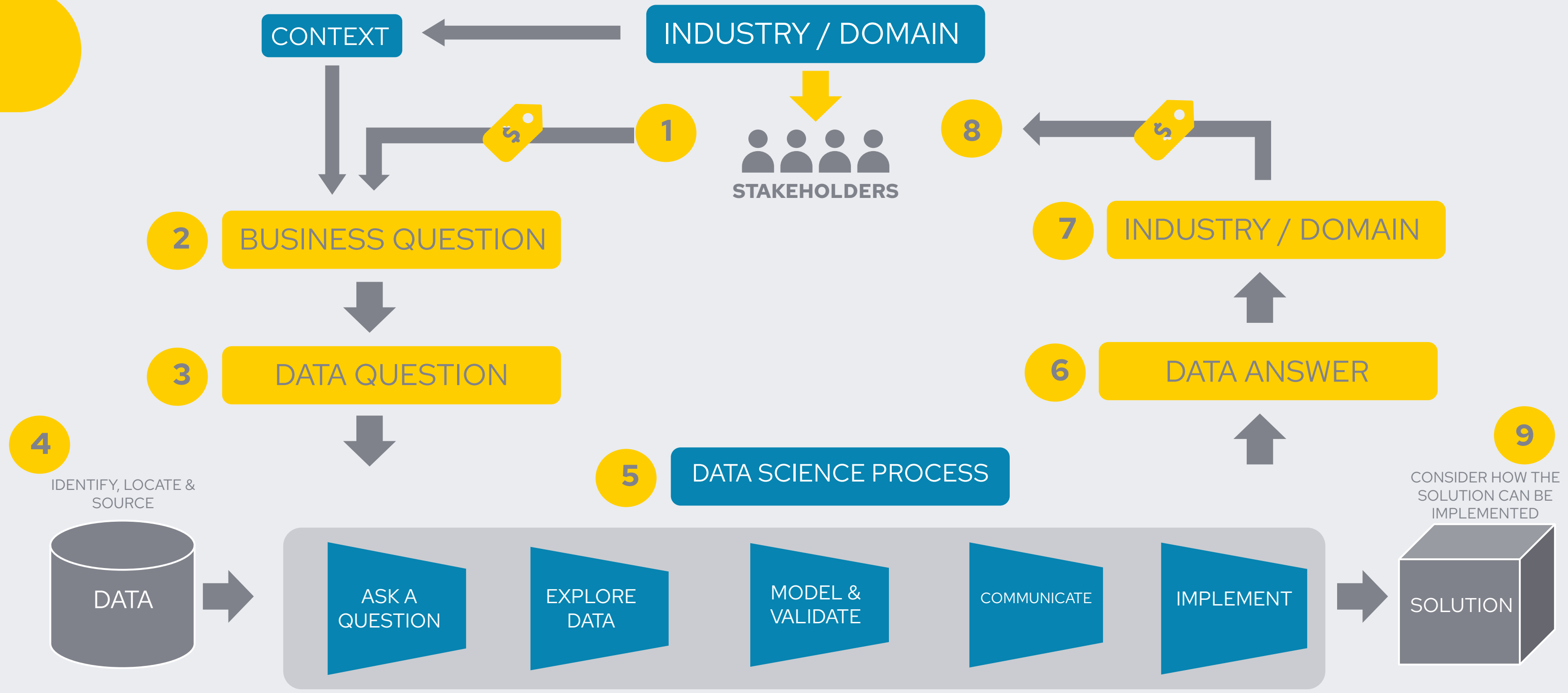
DATASET FEATURES OVERVIEW

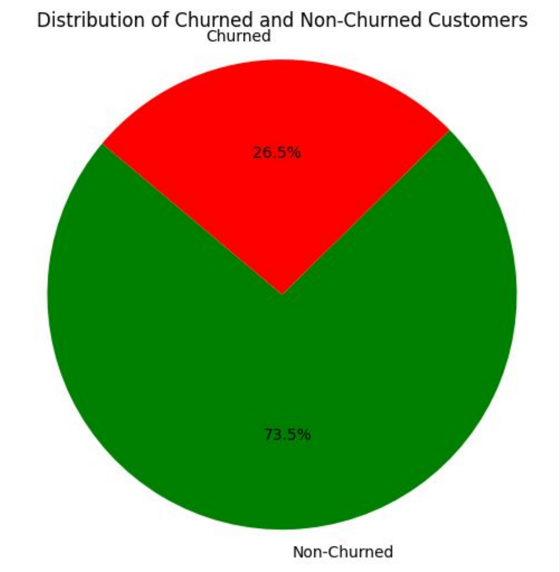
COLUMN NAME	DESCRIPTION	DATA TYPE
customer_id	Unique identifier for each customer	Nominal
gender	Gender	Categorical (M/F)
senior_citizen	65 or older	Binary (Yes/No)
partner	Married	Binary (Yes/No)
dependents	Lives with any dependents	Binary (Yes/No)
tenure	Total months w/ company	Numeric
phone_service	home phone service	Binary (Yes/No)
multiple_lines	Multiple telephone lines	Binary (Yes/No)
internet_service	Internet service type	Categorical (No, DSL, Fiber Optic)
online_security	Additional online security service	Binary (Yes/No)
online_backup	Additional online backup service	Binary (Yes/No)
device_protection	Device protection plan	Binary (Yes/No)
tech_support	Technical support plan with reduced wait times	Binary (Yes/No)
streaming_tv	Television streaming	Binary (Yes/No)
streaming_movies	Movie streaming	Binary (Yes/No)
contract	Contract type	Categorical (Month-to-Month, One Year, Two Year)
paperless_billing	Paperless billing	Binary (Yes/No)
payment_method	Bill payment method	Categorical (Bank Withdrawal, Credit Card, Mailed Check)
monthly_charges	Total monthly charges (all services)	Numeric
total_charges	Total charges, calculated to the end of the specified quarter.	Numeric
churn	Status of the customer at the end of the quarter, Churned or Stayed.	Binary (Yes/No)

3 DESIGN

EDA - Exploratory Data Analysis

APPLYING DATA SCIENCE IN AN INDUSTRY PROJECT





Data Imbalance
- More loyal SpiderCom customers than churners

Modeling Challenge
- Risk of overlooking churners in predictions

Churn Insight Importance
- Fundamental for SpiderCom's operations

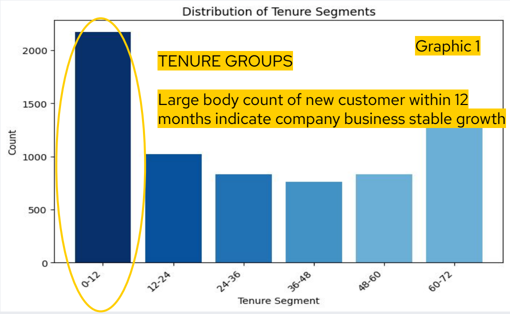
Metrics Focus
- Not just accuracy
- Emphasise precision and recall

Addressing Imbalance
- Explore resampling
- Consider tailored algorithms

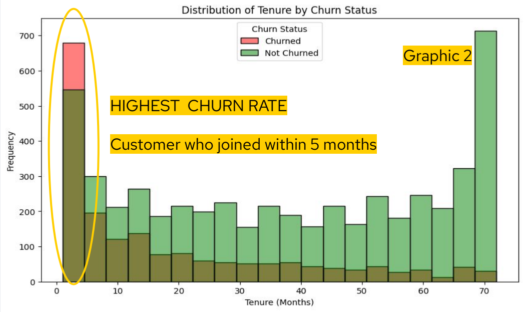
Key factors that influence customer churn rates:

- Tenure
- Contract type
- Payment method
- Monthly charges and total charges
- Gender
- Partner and dependent status
- Senior citizen

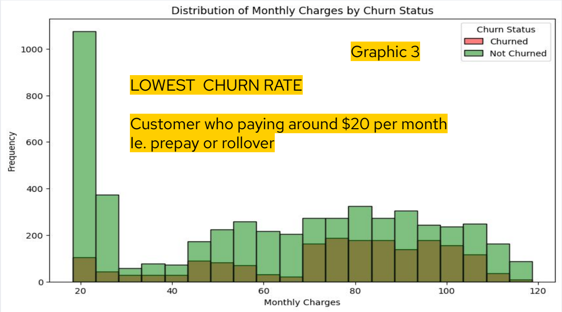
TENURE GROUPS

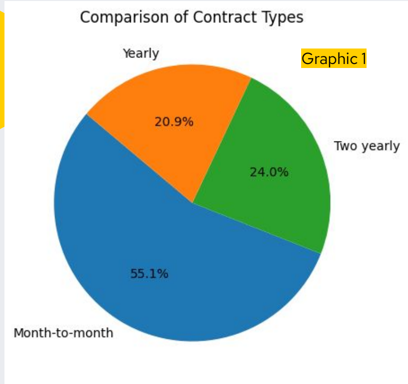


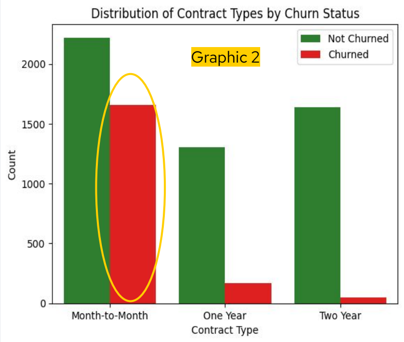
CHURN RATE BY TENURE

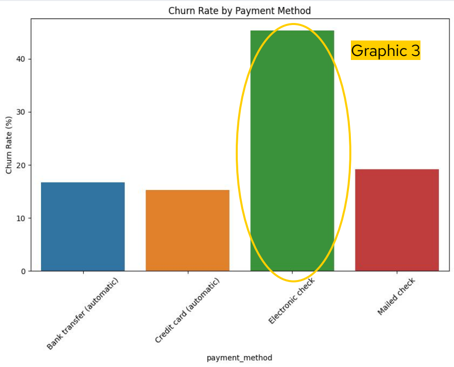


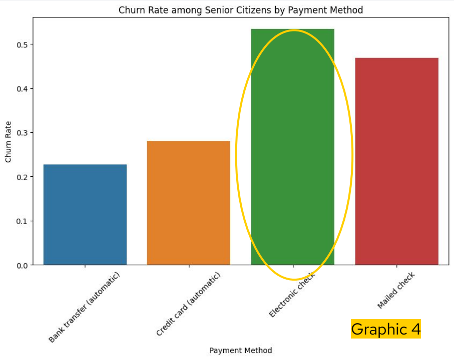
MONTHLY CHARGES

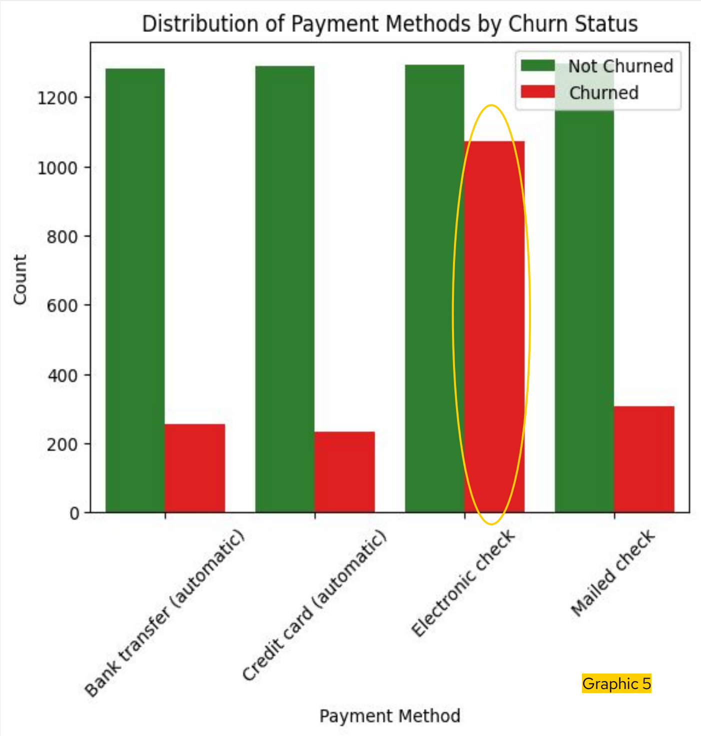








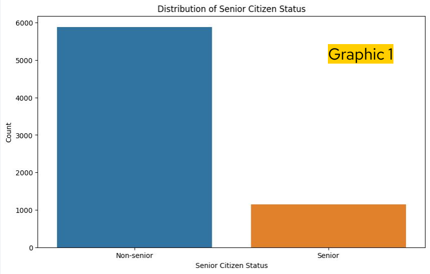




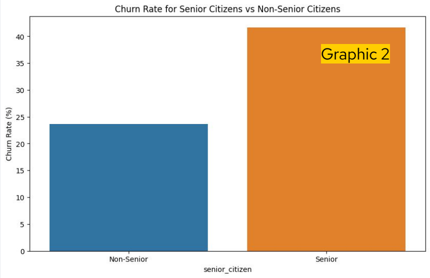
HIGHEST CHURN RATE
Customers not bonded by contract

HIGHEST CHURN RATE BY PAYMENT METHOD
Customers paying by Echeck

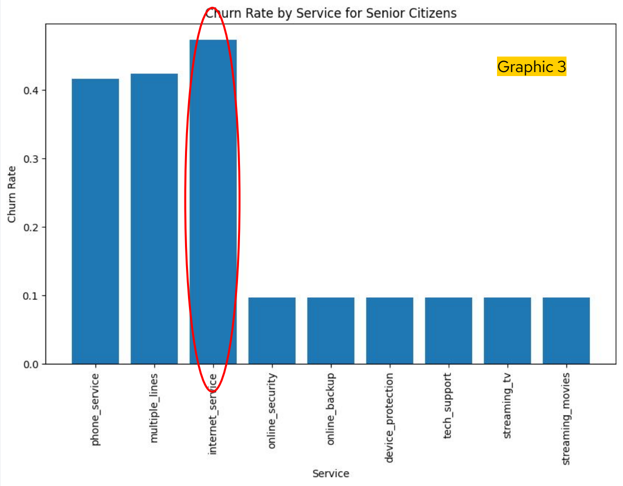
SENIOR CITIZENS



CUSTOMER SEGMENT
<65 GROUP VS SENIOR CITIZENS



HIGH CHURN RATE
IN SENIOR CITIZENS

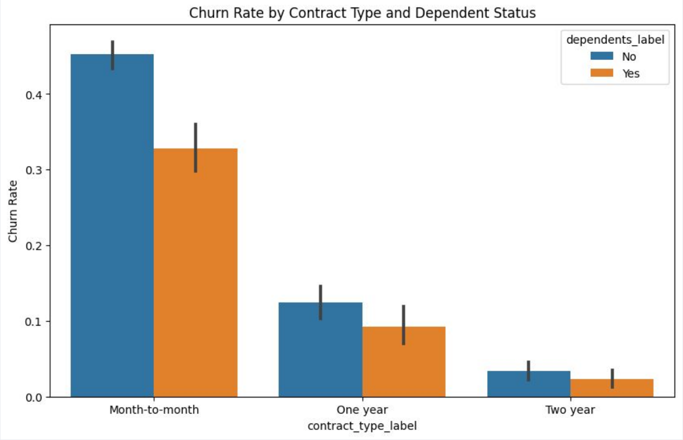


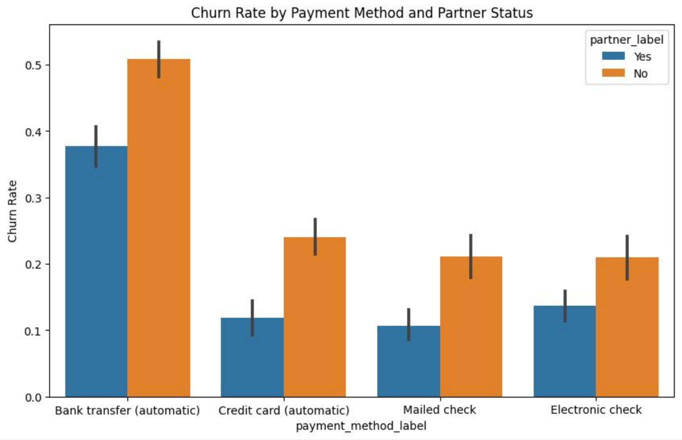
Senior citizens (>65yo) make up a relatively low percentage of the data pool. They have different needs and services than a younger person

- **Needs special assistance:** Senior citizens have unique needs and if not met, they may churn.
- **Price Sensitivity:** Seniors might be more sensitive to pricing and perceived value, leading to higher churn rates (pension).
- **Technology Adoption:** Difficulties in technology use can cause higher churn rates among seniors.
- **Customer Service:** Poor customer service impacts seniors more, contributing to increased churn.

PARTNER / DEPENDENTS

Churned customers		
Partner	**Dependents**	**%**
No	No	34.2
Yes	Yes	21.5
Yes	No	25.4
Yes	Yes	14.3





4 DELIVERY

Feature Engineering & Machine Models

DELIVERY
FEATURE ENGINEERING

Column name	Description
senior_tech_support	Senior customer's tech support status.
tenure_segment	Customer's service length category.
age_segment	Age group category.
total_services	Count of subscribed services.
average_monthly_charge	Average monthly fee.
contract_tenure_interaction	Contract and tenure relationship.
tenure_bin	Binned service length.
monthly_charges_squared	Square of monthly fees.
total_online_services	Count of online subscriptions.
has_streaming	Streaming service status.
log_total_charges	Log-transformed total charges.

ADDED COLUMNS

Creating/Transforming Features:
Senior Tech Support: Combine senior status & tech support.
Log Total Charges: Log transform for nonlinear patterns.

Important Features:
Total Online Services: Sum of online services for engagement.
Tenure Segment: Categorize tenure for loyalty insights.
Has Streaming: Binary encoding for streaming services.

Patterns or Trends:
Senior Tech Support and Churn: Less churn using tech support.
Monthly Charges Squared: Nonlinear pattern in spending & churn.

GRADIENT BOOSTING

```
Accuracy (Gradient Boosting): 0.75
Confusion Matrix:
[[752 282]
 [ 68 306]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.73      0.81      1034
           1       0.52      0.82      0.64       374
```

- Handles unbalanced data; captures complex relationships.
- Robust to overfitting; provides feature importance.
- Scalable; high predictive accuracy.

RANDOM FORESTS

```
Accuracy: 0.75
Confusion Matrix:
[[749 285]
 [ 72 302]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.72      0.81      1034
           1       0.51      0.81      0.63       374
```

- Handles unbalanced data; captures complex dependencies.
- Robust to overfitting; parallelizable; high accuracy.
- Efficient with high dimensionality.
- SMOTE (Synthetic Minority Over Sampling) represent the minority class

LOGISTIC REGRESSION

```
Accuracy: 0.74
Confusion Matrix:
[[744 290]
 [ 72 302]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.72      0.80      1034
           1       0.51      0.81      0.63       374
```

- Interpretability; efficient for large datasets.
- Regularization against overfitting; well-suited for linear relationships.
- Few hyperparameters; applicability in churn.

ENSEMBLE STACKING (ALL 3 MODELS)

- Enhanced accuracy; reduces overfitting.
- Robust & flexible; handles imbalanced data.
- Tailors to business needs; comprehensive churn solution.

DELIVERY
MACHINE MODELS

Data splitting:
- 80/20 split (endures a robust model training & validation)

Model selection:
- (Techniques like Gradient Boosting, Random Forests and Logistic regression classification were used) - Feature selection using multivariate analysis, random undersampling for imbalance

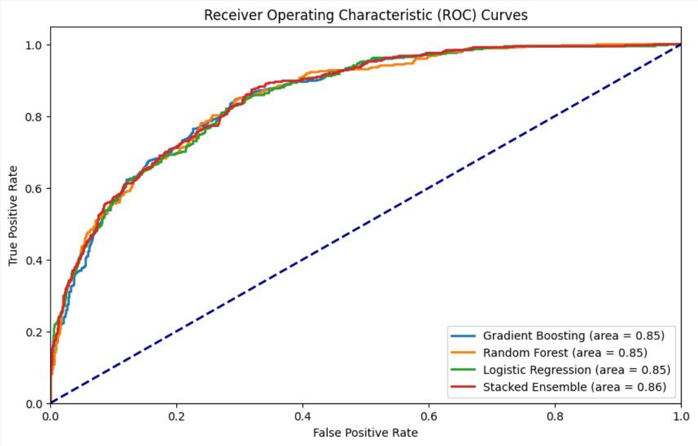
Model training:
- Utilised specific hyperparameters and cross-validation

Model evaluation:
- Employed various metrics to gauge effectiveness of models performance
- (accuracy, precision, recall, F-1 and AUC-ROC).

Hyperparameter tuning (optimise performance)
- Enhanced performance through optimisation (Gridsearch)

Integration:
- All elements were aligned for precise customer churn prediction

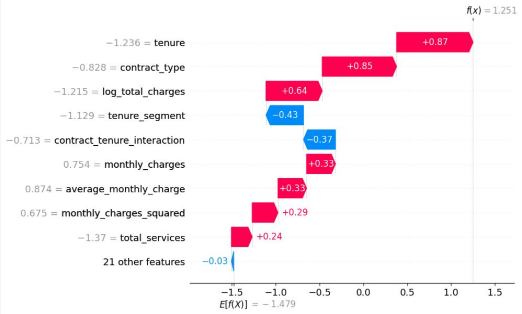
STACKING ENSEMBLE – ROC-AUC



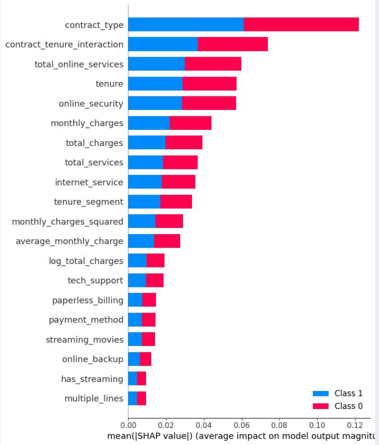
STACKING ENSEMBLE – PERFORMANCE

```
Accuracy (Stacking Ensemble): 0.75
Confusion Matrix:
[[755 279]
 [ 74 300]]
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.73      0.81      1034
           1       0.52      0.80      0.63       374

    accuracy                           0.75      1408
   macro avg       0.71      0.77      0.72      1408
weighted avg       0.81      0.75      0.76      1408
```

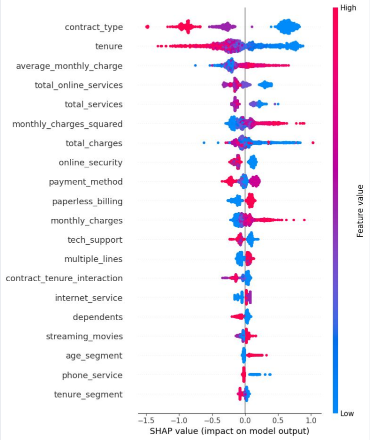
LOGISTIC REGRESSION (SHAP)



RANDOM FORESTS (SHAP)



GRADIENT BOOSTING (SHAP)



5

SUMMARY , CONCLUSION & NEXT STEPS

SUMMARY & CONCLUSION

Business Question:

- How do we reduce customer churn and enhance loyalty at Spidercom?

Key Insights:

- Churn Influencers: Tenure, Contract, Monthly Charges, Payment Method.
- Demographic Factors: Gender, Partner/Dependent status, Senior Citizen.

Model Performance:

- Accuracy: 75% in predicting churn.
- Consistency: cross-validation accuracy.

Strategic Achievements:

- Insightful Analysis: Identified key factors affecting churn.
- Robust Models: Including Stacking Ensemble for predictive insights.
- SHAP Analysis: Transparent evaluation of feature influence.

Actionable Conclusion:

- Targeted strategies based on key influencers can enhance loyalty.
- Opportunity for tailored offerings and personalized customer engagement.

NEXT STEPS

APP DEPLOYMENT

- Model deployment: application
  CRM, Marketing team

CRM: Customer Relationship Management
(Decision support system that manages the interactions between an organisation and its customers)

Database marketing: using CRM databases to develop one-to-one relationships and precisely targeted promotional efforts with individual customers (3).

NEXT STEPS

Future steps: potential improvements to the model, expanding the application's functionalities, or exploring other data science use cases within the telecommunications industry.

ACKNOWLEDGMENTS

<u>Institute of Data:</u>
IOD for providing me with the opportunity to undertake this Data Science & Artificial Intelligence qualification and capstone project. The knowledge and skills gained during this journey have been invaluable for my professional growth.

<u>Trainers:</u>
Amin, Sakshi, Isabelle, for their guidance, mentorship, and valuable feedback throughout the course. Their expertise and support have been instrumental in shaping the success of this endeavor.

6 QUESTIONS

7 APPENDIX

Reference Documents

APPENDIX

1. Hussain, S. (2016). Bankers, Hug Your Customers: A Guide to Every Banker to Delight Customers, Employees, and Colleagues. United Kingdom: Partridge Publishing India.

2. Dahl, J. (2019). Leading Lean: Ensuring Success and Developing a Framework for Leadership. Taiwan: O'Reilly Media.

3. Zikmund, W. G., Ward, S., Lowe, B., & Winzar, H. (2007). Marketing Research (8th ed.). South Melbourne, Australia: Cengage Learning Australia.

Data source: Open ML https://api.openml.org/d/42178