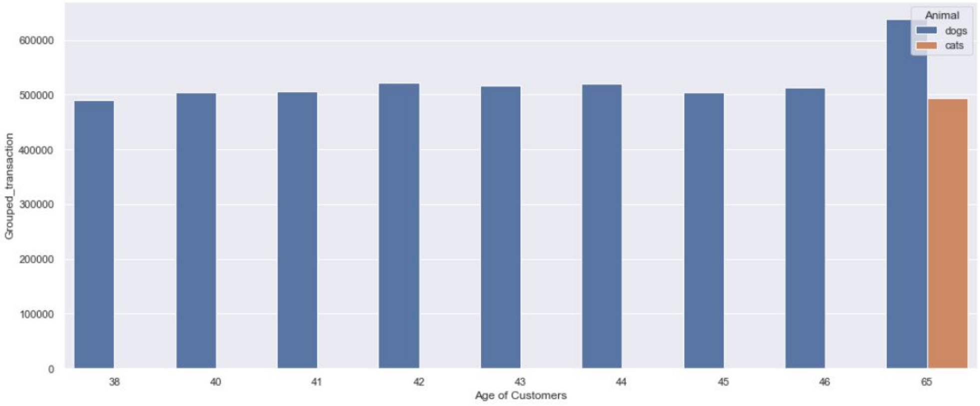
Fressnapf

Expert Data Scientist
Case Study

Top 10 **profitable** products by **Animal**, **Category_type** and **Age of Customers**

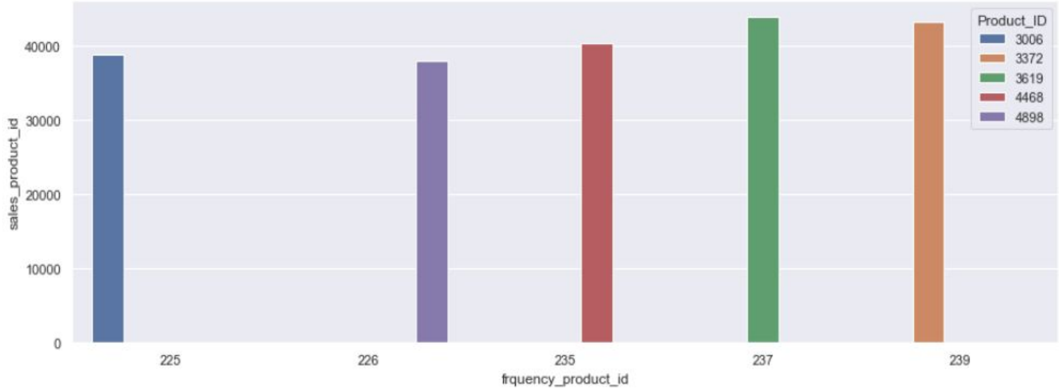
Definition of Profitable = Unit Cost (per product) * Purchase Quantity (per customer)



	Animal	Category_type	Age of Customers	Grouped_transaction
0	dogs	food	65	638223.88
1	dogs	food	42	522515.50
2	dogs	food	44	519736.03
3	dogs	food	43	516294.99
4	dogs	food	46	513155.57
5	dogs	food	41	505727.84
6	dogs	food	45	504036.62
7	dogs	food	40	503764.04
8	cats	food	65	493264.91
9	dogs	food	38	490517.48

The top 10 contributors to Fressnapf's profit are customers over the age of 35 purchasing products from the food category for their pets (Dog).

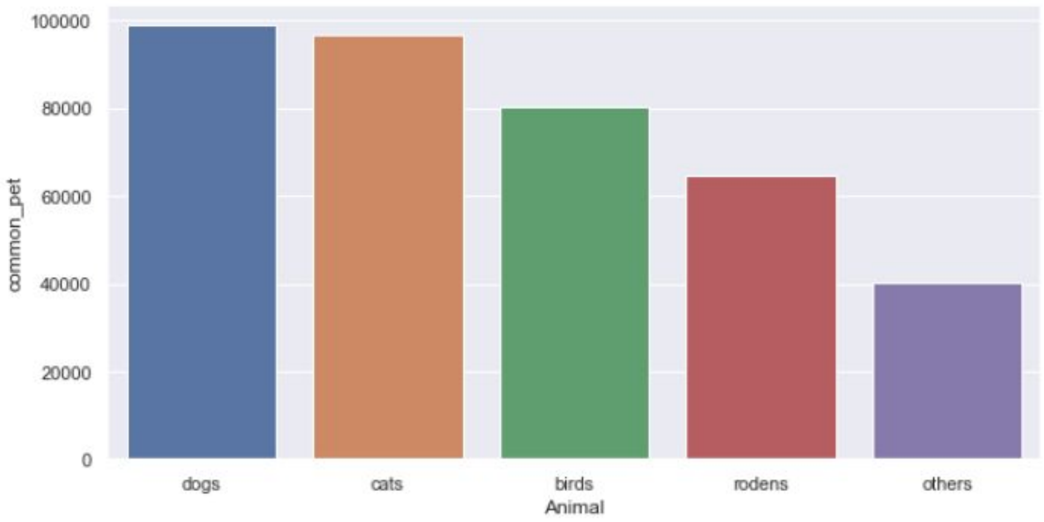
Profitable Vs Frequent



	Product_ID	frquency_product_id	sales_product_id
3519	3619	237	43907.50
3272	3372	239	43362.20
4368	4468	235	40484.82
2906	3006	225	38862.74
4798	4898	226	38080.88

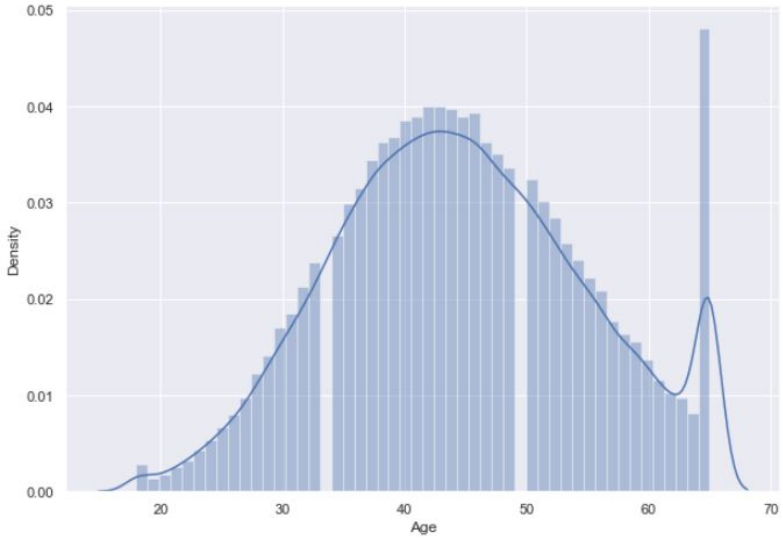
- **Product profitability** is not significantly different from **product frequency**.
- From the green (Product_ID: 3619) and brown (Product_ID: 3372) we can conclusively say most profitable product is the most frequent one and vice versa.

Most Common Pet



Customer Segment

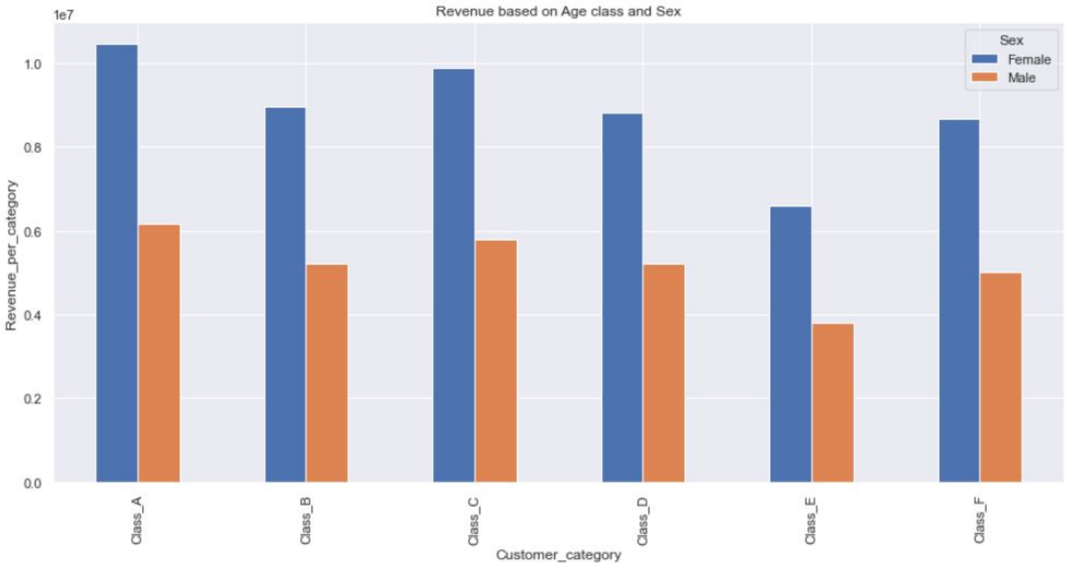
Research Question: From the distribution of customers age, can we predict the age group of our customers given the details of their purchase history.



Profitable Customer Segment

Customer Segments (chosen intervals are based on the distribution)

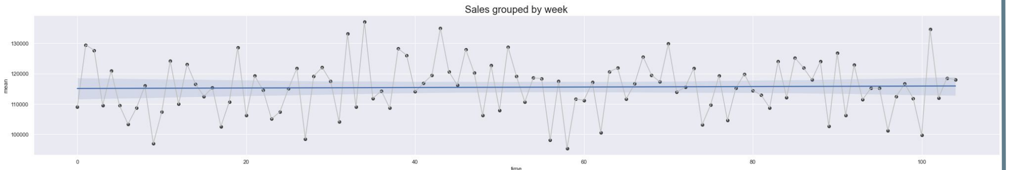
- Class_A - Age below 35 Years
- Class_B - Age between 35 - 40 Years
- Class_C - Age between 40 - 45 Years
- Class_D - Age between 45 - 50 Years
- Class_E - Age between 50 - 55 Years
- Class_F - Above 55 Years

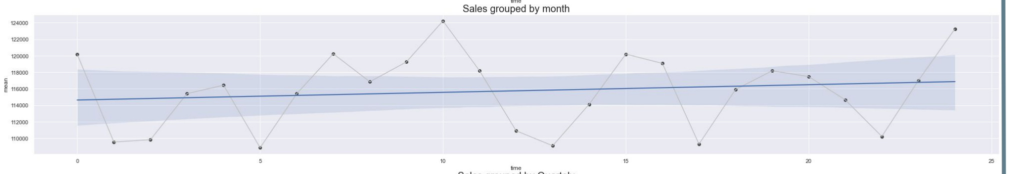


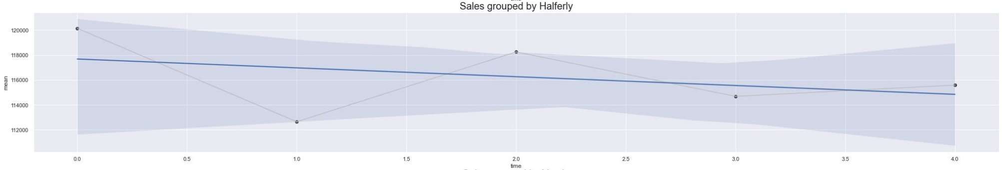
Female and **Male** customers with age **below 35** are the **most profitable** customer group for Fressnapf

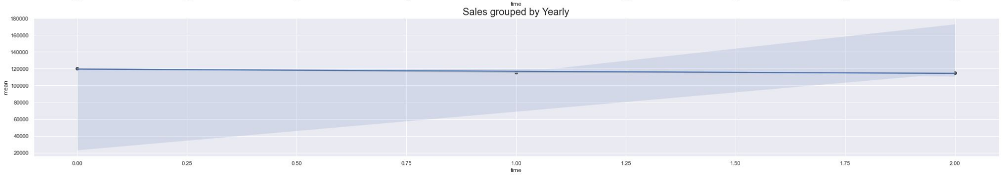
Trend Determination

Grouped Revenue: Trend (Blue line), Rolling mean (Grey line).

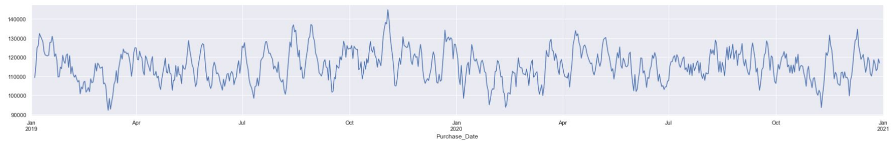






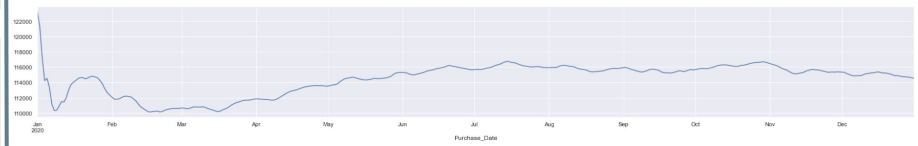


Trend component 2019 - 2020



Trend component 2019

Trend component 2020



Churn Rate

- Churn Rate - Expressed as percentage of customers stopped purchasing from Fressnapf after a specific time period.
- Churn rate = (number of lost customers / total amount of customers) * 100 *
- For example if fressnapf loses 10 customers from a 1500 customer base, the churn rate would be (10/1500)*100 = 66%.
- Actual churn rate from the given dataset

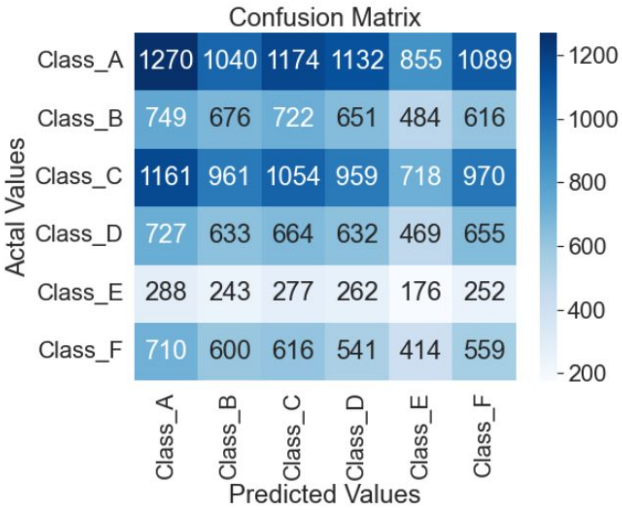
Year and Window	Churn rate
Y - 2019, W - 90 days	23.47 %
Y - 2020, W - 90 days	25.49 %

Model Development - Customer Segmentation

1. Segmentation Types
    a. Target - 1: Customers are grouped by the Age group (Class_A to Class_F)
    b. Target - 2: Customers are grouped by the gender (Male, Female)
2. Feature Engineering
    a. Number of Transactions per customer for the period 2019 - 2020 (MinMax Scaling)
    b. Transaction amount per customer (Power Transformation)
    c. Number of Transactions per customer per channel (MinMax Scaling)
    d. Transaction amount per customer per channel (Power Transformation)
    e. Category encoding of product Category_type
    f. Category encoding of product Animal group
    g. Day of the week (Monday - Sunday) encoding of Purchase_date
    h. Standard deviation of purchase interval per customer (MinMax Scaling)
    i. Average purchase quantity per transaction (MinMax Scaling)

Results: Target - 1 – Age Group (Class_A to Class_F)

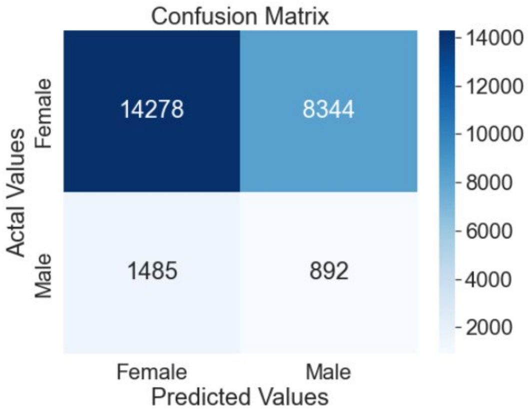
- Accuracy - 17.5 %
- F1 - Score (Macro Avg) - 16 %
- Confusion Matrix



- Reason - Underfitting. The features provided cannot be used for segmenting the customers by their Age group

Results: Target - 2 – Gender (Male and Female)

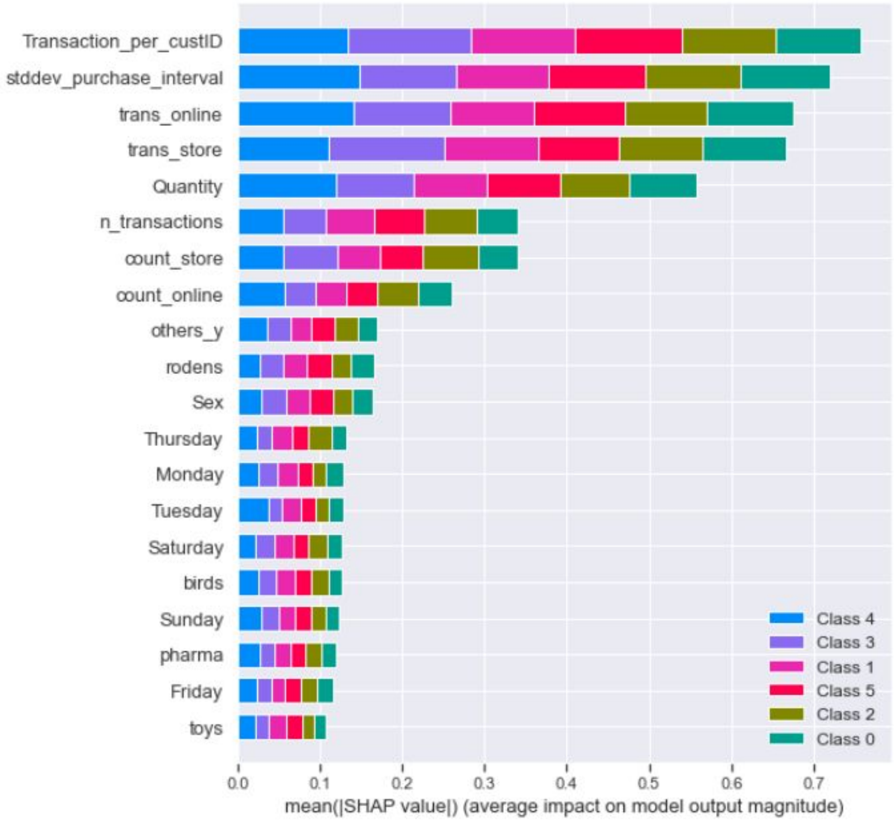
- Accuracy - 61 %
- F1 - Score (Weighted Avg) - 70 %
- Confusion Matrix



- Reason - Partially underfitting. The generated features have high correlation between them.

- More features that distinctly represent the gender of the customer has to explored.

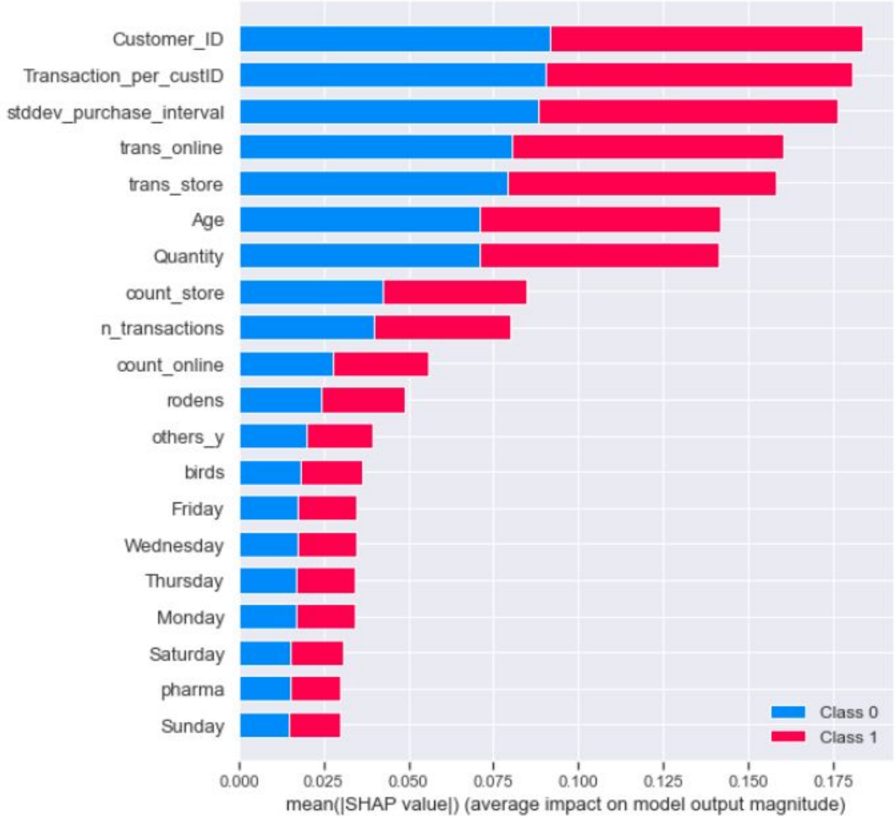
Feature Importance and Contribution - Target 1 – Age Group Segmentation



- Food category, Animal group and Purchase day of the week are insignificant and less utilized by the model.
- Top 8 features share the impact (SHAP value) on the model output (Class) equally.
- Features specific to each class has to be explored for increasing the model performance.

Note - For SHAP summary plot per class and per sample check the notebook.

Feature Importance and Contribution - Target 2 – Gender Segmentation



- Except Purchase day of the week all other drafted features are significant.
- Positive: Differentiation of the boundaries between gender class is different for each feature.
- More customer data should be converted to usable features which will consequently improve the model performance.

Note - For SHAP summary plot per class and per sample check the notebook.

Project Value to Fressnapf

- The **Customer Segmentation** on **Gender** shows noteworthy results which can be further pursued to improvise the performance of the machine learning algorithm.

- The output of a similar project on segmenting customer on different basis can be utilized to:
    - Dynamic strategies and campaigns for the dominant customer group
    - Increase ROAS by directing the marketing funding to the specific group
    - Behaviour analysis of different customer group
    - Personalized product recommendation based on the group
    - Higher customer retention and reduction in churn rate therefore higher revenue generation
    - Directed and precise customer acquisition

Cloud Infrastructure Requirement

Problems

- No enough CPU cores to parallelize the GridSearch and Bayesian Optimization algorithm.
- Slower computation of SHAP values due to large dataset of > 90000 rows. Exponential increase in computation time with increase in drafted features.
- ELT (Extract, Load and Transform) operations are computationally expensive with limited memory and processing power.

Cloud Infrastructure Requirement

- Minimum infrastructure requirement – Certain selection criteria are to be considered based
  - Access pattern, Access type, File size, Concurrency
- Data collection - Distributed cloud infrastructure (eg: BigQuery) for querying and processing of customer data and product meta data.
- Data preparation (Anomaly detection, data normalization and labelling) - General purpose CPUs (E2 series, 2-32 vCPUs)
- Quick testing and development (AI platform) - C2 compute-optimized
- ML, DL training and inference - HPC (eg: n1-highmem-16 (16 vCPUs, 104 GB RAM))

Productionization Requirements for Customer Segmentation

- Model Serving
  - *Online prediction service* – Kubernetes Cluster with minimum 6 vCPU and maximum 10 vCPU, memory of 8 to 12 GB, compute optimized C2 machine series, autoscaled nodes from minimum 4 to maximum 10. Persistent volume of 1 GiB.
  - *Offline batch processing* – Virtual machines with 4 to 8 vCPU with memory of 8 GB.
- Feature store - The customer data is prone to drift based on the changes in purchase behaviour therefore features have to be computed and stored for the iterative model training.
- Docker, Artifacts, Model and Source code registry on cloud environment for performance tracking and monitoring.
- Kubeflow or mlflow for MLOps process.

THANK YOU