Predictive Models and Chatbot

IS483 Final Presentation

Data Ninjas (Team ID23)
Darren Png | Neo Jia Ying
Nor Aisyah Bte Ajit | Tay Yu Liang
Wong Wei Ling | Yeo Hui Xin

SINGAPORE MANAGEMENT UNIVERSITY

PRUDENTIAL

TABLE OF CONTENTS

INTRODUCTION

Client, Group &
Project Overview

OUR CLIENT

- Prudential Assurance Company Singapore
- One of Singapore's leading life insurance companies
- Products offered include Life, Health and Wealth Insurance

Main Point of Contact

- Innovation Department
- Magdalene Loh - Head of Innovation
- Luis Aw - Innovation Executive

Stakeholders

- Digital Content Team
  - Alice Yu - Digital Content Lead
- OPEX (Operational Excellence) Team
  - Madhan Seduraman - Head of OPEX
  - Zhang Siqi - Chatbot Engineer

OUR TEAM

Darren Png
Data Analyst / Quality Assurance

Yeo Hui Xin
Data Analyst / Quality Assurance

Neo Jia Ying
Data Analyst / Project Manager

Tay Yu Liang
Data Analyst / Coordinator

Wong Wei Ling
Data Analyst / Secretary

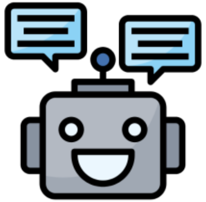
Nor Aisyah Bte Ajit
Data Analyst / Secretary

PROJECT SCOPE

Improving The Customer (Gen Y and Z) Journey In Insurance

Customer Acquisition through Targeted Marketing

- Find out the opinions of Gen Y and Z on Insurance and ILPs
- Understand the needs of different segments of Gen Y and Z

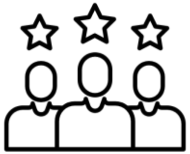


Enhancing Customers Engagement using IBM Watson Chatbot

- A FAQ answering chatbot to improve customers' experience
- Topic modelling incorporated for commonly asked questions regarding Insurance from online sources

Customer Risk Assessment using Machine Learning Models

- An alternative to cross check the assessment of customer risk
- Makes the process quicker and less labor intensive for new customers to get a premium



Increasing Customer Retention using Machine Learning and Clustering models

- Understand why customers churn or stay
- Come up with targeted solutions to improve customer loyalty of different segments of customers

PROJECT MOTIVATIONS

1. Venture into their **desired market segment** (Gen Y and Z - **21 to 36 years old**)
    - Develop **customized marketing campaigns and policies** to attract more customers from their targeted market
2. Enhance **customer service**
    - Needs to know the FAQs customers have and incorporate these topics using topic modelling into chatbot
3. Increase **customer satisfaction**
    - Automate risk assessment and speed up insurance purchase process
4. Improve **customer loyalty** to increase profitability
    - Targeted solutions for each customer segment

USER JOURNEY

Stages

Activities

Feelings

Targeted Marketing

Alice:

1. Promote a policy to Gen Y and Z
2. Learn what Gen Y and Gen Z are interested in
3. Discover the different customer segments
4. Explores the dashboard on their opinions on Insurance and ILPs
5. Develop targeted marketing campaign(s) for Gen Y and Z

"I know what Gen Y and Z look out for when buying insurance policies"

Enhance Customers' Engagement

After increasing customers' awareness through Alice's marketing campaign…

Siqi:

6. Deploys the chatbot onto prudential website to handle the influx of common questions asked
7. Interested customers have their questions answered immediately
8. Customers request for a purchase.

"Now I can attend to these potential customers' queries more efficiently"

Customer Risk Assessment

After receiving the purchase request…

Madhan:

9. Input customer information into risk assessment predictive model
10. Adjust premiums accordingly to customer's risk level.
11. Provide quotation of the premium to customer
12. Purchasing process ends with a satisfied customer.

"I can tailor to specific customer needs by coming up with customised premiums!"

Customer Retention

When customer premium is reaching maturity….

Madhan:

13. Uses retention predictive model to predict customer churn rate
14. With the clustering model, identify the reasons why customer exit or stay
15. Identify factors and improve on increasing customer loyalty.

"I can finally increase the number of happy and loyal customers!"

USE CASES

Use Case for Customer Acquisition through Targeted Marketing

Alice

Persona

- Digital Content Lead
- Digital marketing space
- Ideate & advocate Prudential's Products using social media platforms

Expectations

- To understand Gen Y/Z **perception** on **insurance** and **ILPs** using **text mining** techniques and **clustering**

Opportunities

- To craft **targeted marketing strategies** and appeal to Gen Y/Z
- To **cross sell similar products** to Gen Y/Z to increase Prudential's sales

USE CASES

Use Case for Enhancing Customers Engagement using IBM Watson Chatbot

Si Qi

Persona

- Team Lead for Prudential's Operational Excellence (OPEX) Department
- Enhance Prudential's customer satisfaction
- Reduce the company's operational inefficiencies

Expectations

- To **identify** customers **queries** and implement a chatbot to **automate** frequently asked questions.

Opportunities

- To enhance the **intuitiveness of existing chatbot** using information obtained from the general public
- Aims to provide **commonly asked questions** and **improve customers' satisfaction**.

USE CASES

Use Case for Customer Risk Management using Machine Learning Models

Madhan

Persona

- Head of Prudential Operational Excellence (OPEX) Department
- Focuses on enhancing Prudential’s operations

Expectations

- To **identify** which customers have **high risk**,
- To **efficiently** adjust premiums accordingly to the evaluations

Opportunities

- Further understand customers
- Identify customers’ characteristics that account for their **risk value**
- To **reduce potential loss** for Prudential

USE CASES

Use Case for Increasing Customer Retention using Machine Learning and Clustering models

Madhan

Persona

- Head of Prudential Operational Excellence (OPEX) Department
- Focuses on enhancing Prudential's operations

Expectations

- Accurately **predict** customer's **turnover rate**
- To identify **clusters** of customers with similar attributes who are planning to churn

Opportunities

- Further **understand** customers' **needs**
- **Identify** common customers' **characteristics**
- Find out ways on **how to retain customers**

TOOLS USED

- Data Preprocessing
- Natural Language Processing
- Clustering
- Machine learning

- Output csv files

- Chatbot
- Visualization
- Jupyter Notebooks

TECHNIQUES USED

Customer acquisition through targeted marketing

- Natural Language Processing
  - Lemmatization
  - Tokenization
  - Stopwords removal
  - POS Tagging
  - CountVectorizer
  - Topic Modelling
  - LDA
- K-means clustering

Enhancing Customers Engagement using IBM Watson Chatbot

- Natural Language Processing
  - Lemmatization
  - Tokenization
  - Stopwords removal
  - POS Tagging
  - CountVectorizer
  - Topic Modelling
  - LDA
- IBM Watsons

Customer Risk Assessment using Machine Learning Models

- Machine Learning
  - Logistic Regression
  - Random Forest
  - Decision Tree
  - KNN
  - Adaboost
  - XGBoost Classifier
- Feature selection (RFECV)
- SMOTE
- K Fold Cross Validation
- RandomizedSearchCV

Increasing Customer Retention using Machine Learning and Clustering models

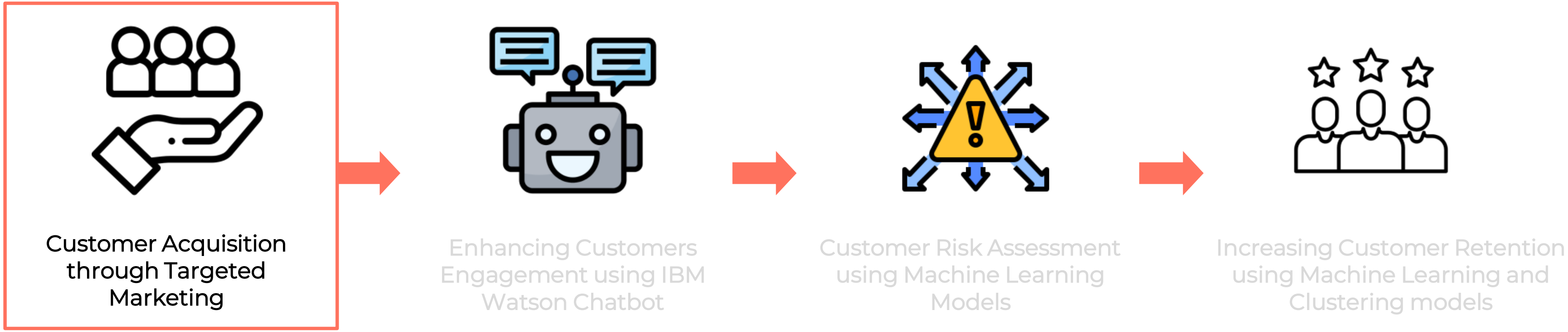
- Machine Learning
  - Random Forest
  - Extra Trees Classifier
  - Super Vector Classifier
  - Adaboost Classifier
  - Gradient Boost Classifier
  - XGBoost Classifier
- SMOTE
- K Fold Cross Validation
- GridSearchCV
- K-means clustering

Customer Acquisition through Targeted Marketing

Gen Y and Z

RECAP: PROJECT SCOPE

Improving The Customer (Gen Y and Z) Journey In Insurance



GOOGLE SURVEY

Opinions on Insurance and ILPs

Solution Architecture

Design Google Forms survey

Export csv files for cleaning & analysis

Clean using **Pandas, Regex, Numpy**

Opinions on Insurance + ILP

Natural Language Processing (NLP) using **NLTK, Matplotlib, Sklearn_lda;**

Export required files for visualization

Power BI

Visualize findings using **PowerBI** & derive recommendations

Opinions on Insurance

Clustering using **Sklearn**

jupyter

Derive insights

GOOGLE SURVEY

<table>
<tr><th>Opinions on Insurance</th><th>Opinions on ILPs</th></tr>
<tr><td colspan="2">

Designed google forms to ask for their age as Prudential is interested in Gen Y and Z

Which one of the following represents your age?
○ 21 years old and under
○ 22 to 36 years old
○ 37 to 51 years old
○ 52 years old and above

Disseminated in SMU Telegram chats (eg. AskSMU, SISters) → Since SMU students belong to Gen Y and Z
CSV exported from Google Forms

</td></tr>
<tr><td>179 Respondents</td><td>144 Respondents</td></tr>
</table>

DATA PREPROCESSING

Opinions on Insurance and ILPs

Convert to lowercase → Remove punctuations → Strip whitespaces

→ Rename columns → Drop timestamp → Extract values

NATURAL LANGUAGE PROCESSING (NLP)

Opinions on Insurance and ILPs

Solution Architecture

Lemmatization using **WordNetLemmatizer** in **NLTK stem**

Tokenization using **RegexpTokenizer** in **NLTK**

Remove stop words using words in **NLTK corpus**

POS Tagging using **NLTK pos_tag**

CountVectorizer using **Sklearn feature extraction text**

Topic Modelling using **LDA** in **sklearn decomposition** (For Opinions on Insurance)

Power BI

Visualize findings using **PowerBI** & derive recommendations

METHODOLOGY - NLP

- A subset of AI that extracts meaning from human language

1. Lemmatization

- Group different forms of words so that they can be analysed together
- E.g. rocks → rock, better → good

2. Tokenization

- Split samples of text into words

3. Stop Words Removal

- E.g. a, an, the

METHODOLOGY - NLP

❖ A subset of AI that extracts meaning from human language

4. Part of Speech (POS) Tagging

- Categorises words into nouns, verbs, adjectives & adverbs etc.
- E.g. She (pronoun) sells (verb) seashells (noun)
- Keep only the nouns to get more valuable insights

5. Count Vectorizer

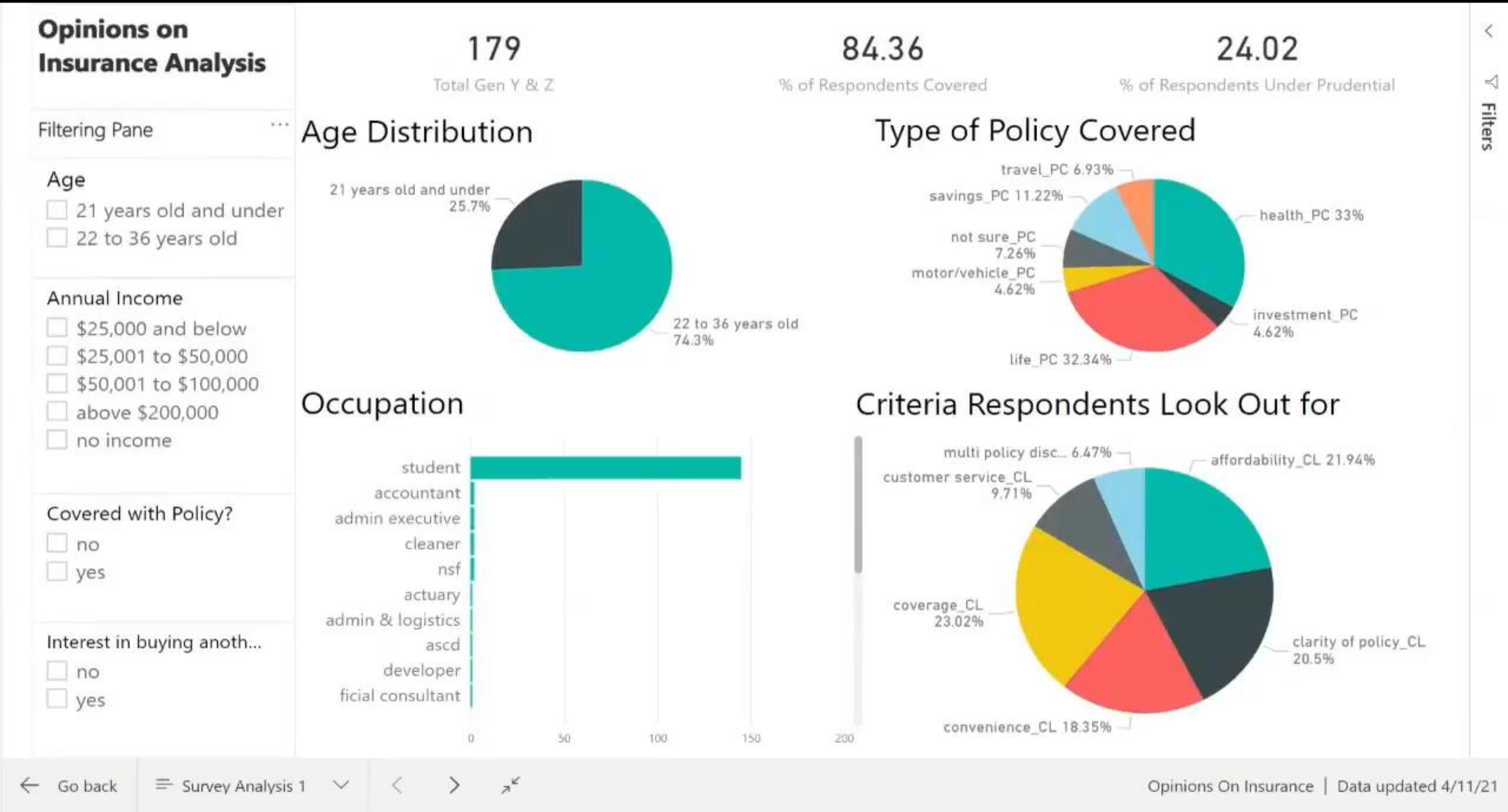
- To get frequency of words

6. Latent Dirichlet Allocation (LDA)

- Unsupervised machine learning model
  - **Document**: a distinct text
  - **Topic**: a group of words from a collection of documents
- **Topic modeling**: Identify topics in a set of documents

Opinions on Insurance

NLP Insights



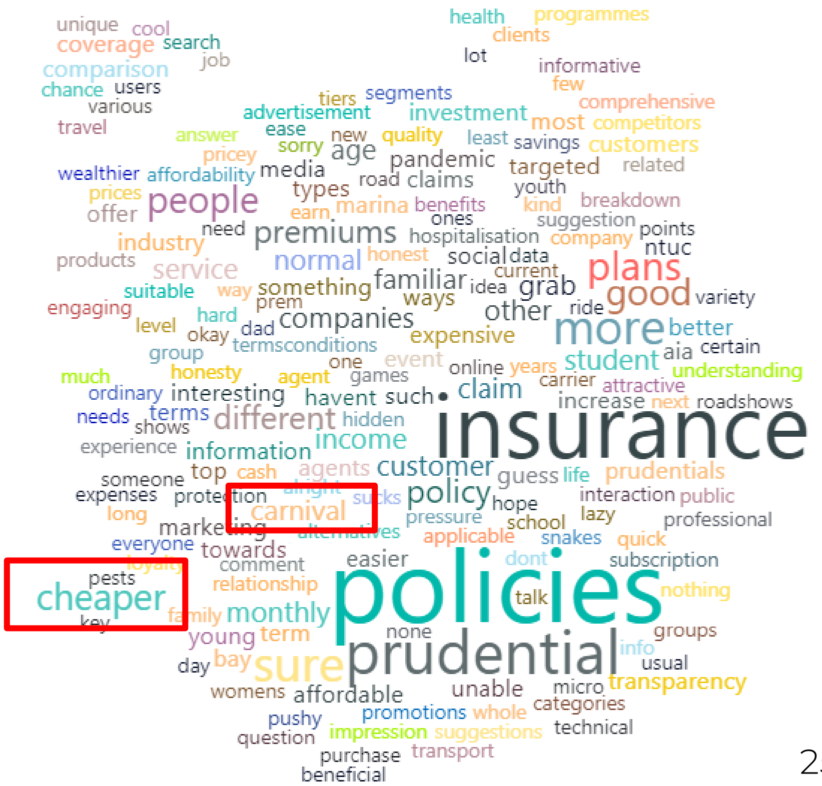
Opinions on Insurance Analysis
179
Total Gen Y & Z
84.36
% of Respondents Covered
24.02
% of Respondents Under Prudential
Filtering Pane
Age
21 years old and under
22 to 36 years old
Annual Income
$25,000 and below
$25,001 to $50,000
$50,001 to $100,000
above $200,000
no income
Covered with Policy?
no
yes
Interest in buying anoth...
no
yes
Age Distribution
21 years old and under 25.7%
22 to 36 years old 74.3%
Type of Policy Covered
travel_PC 6.93%
savings_PC 11.22%
health_PC 33%
not sure_PC 7.26%
motor/vehicle_PC 4.62%
investment_PC 4.62%
life_PC 32.34%
Occupation
student
accountant
admin executive
cleaner
nsf
actuary
admin & logistics
ascd
developer
ficial consultant
0
50
100
150
200
Criteria Respondents Look Out for
multi policy disc... 6.47%
affordability_CL 21.94%
customer service_CL 9.71%
coverage_CL 23.02%
clarity of policy_CL 20.5%
convenience_CL 18.35%
Filters
Go back
Survey Analysis 1
Opinions On Insurance | Data updated 4/11/21

INSIGHTS

Opinions on Insurance

Qn: What do you think of their policies and any improvements/suggestions you hope to see from Prudential?

- **Marketing**:
  - Organise/join school events to enhance brand awareness & reputation e.g. Marina Bay Carnival
- **Additional Perks**:
  - Wealth policies: Offer higher interest rates for Gen Z
- **Cheaper Plans:**
  - Most Gen Y and Z are still schooling, hence they find the plans expensive
  - Suggestions:
    - Tiered plans
    - Monthly subscription plans



INSIGHTS

Opinions on Insurance

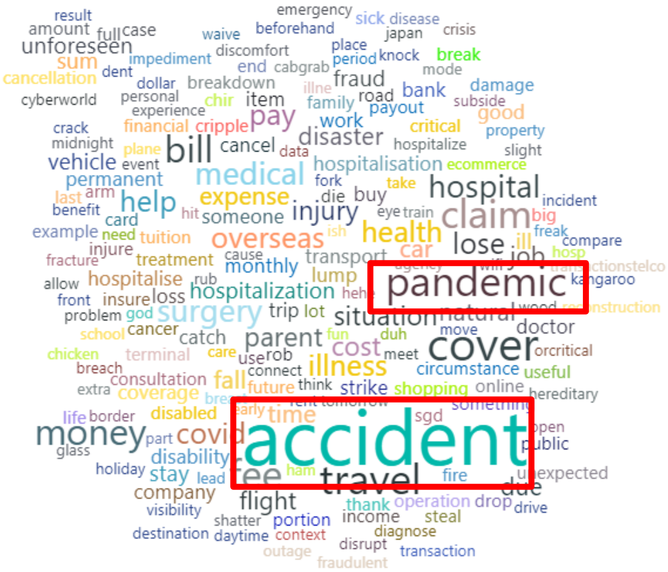
Qn: What do you think of their policies and any improvements/suggestions you hope to see from Prudential?

- **Provide More Information**:
  - More ads on social media platform like Youtube & Instagram to promote policy & its benefits
- **Policies More targeted to Gen Y & Z**:
  - E.g. NTUC Income to cover 60% of riders' Grab, Ryde and GOJEK fare when it rains to shield them from surge pricing when it rains

INSIGHTS

Opinions on Insurance

Qn: Describe a situation in which having an insurance would have helped you.



Topics found via LDA:

Topic #0:
covid pandemic travel claim natural disaster overseas cover hospitalization like

Topic #1:
lose pandemic pay job help far injury fee suddenly transport

Topic #2:
surgery parent situation help stay overseas fall buy hospitalisation hospital

Topic #3:
accident illness cover car pay health medical hit injury loss

Topic #4:
cover accident money help lose fee travel hospital medical cost

- People think that they would benefit from having insurance if:
  - They are in an accident
  - While travelling (when borders are re-opened again)
  - When living in a pandemic (to cover hospitalization & job loss)

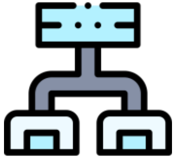
Opinions on Insurance

Clustering

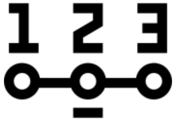
CLUSTERING MODEL

Opinions on Insurance

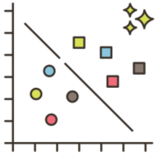
Solution Architecture



Split dataset into **Interested** and **not interested** in buying another policy



Encode categorical columns from string values to numerical values using **LabelEncoder()**



Perform **K-Means Clustering** using **KMeans** in **sklearn cluster** & Cluster **Profiling** by calculating Z-score

Reverse encoding to get back original column values which were string

Plot **Heatmap** to see highly ranked variables in each cluster using **Seaborn**

Compile **insights** for each cluster on **Jupyter**

METHODOLOGY

1) Split dataset

```
#separate df into people who are interested/not interested in buying an insurance
df_interested = df[df['Interest in buying another policy'] == 'Yes']
df_interested.head()
```

2) Label encoding using LabelEncoder()

a) Machine learning algorithms generally support numerical values

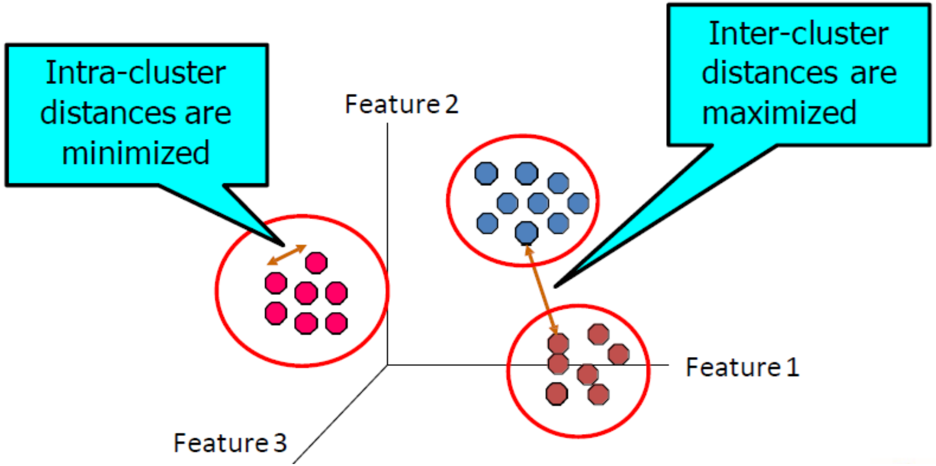
b) Hence, need to convert text data to numeric to perform clustering

Original value	After encoding
21 years old and under	0
22 to 36 years old	1

CLUSTERING

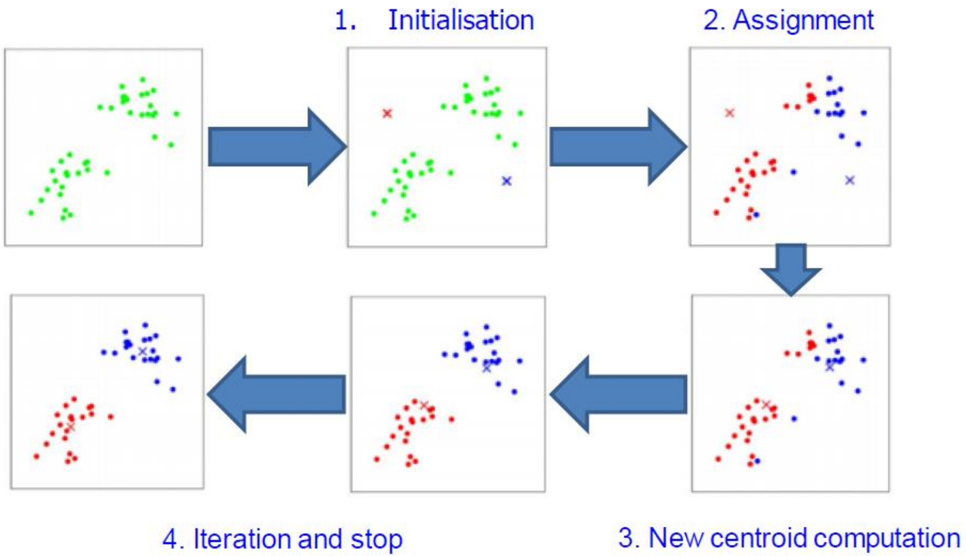
Purpose:

- Finding groups of objects such that
  - Objects **in a group** will be **similar** (or related) to one another (homogeneous)
  - But **different** from the objects **across the other groups** (heterogeneous)

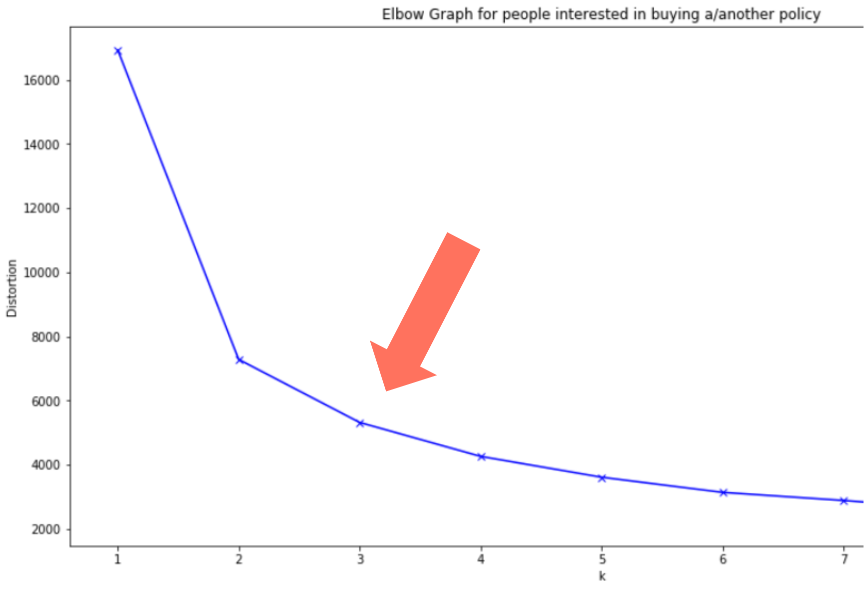


K-MEANS CLUSTERING

- **Partitional clustering**
  a. A division data objects into non-overlapping subsets (clusters) such that each data **object** is in **exactly one subset**
- Uses **Error Sum of Squares (SSE)** to differentiate the clusters



METHODOLOGY – CLUSTERING

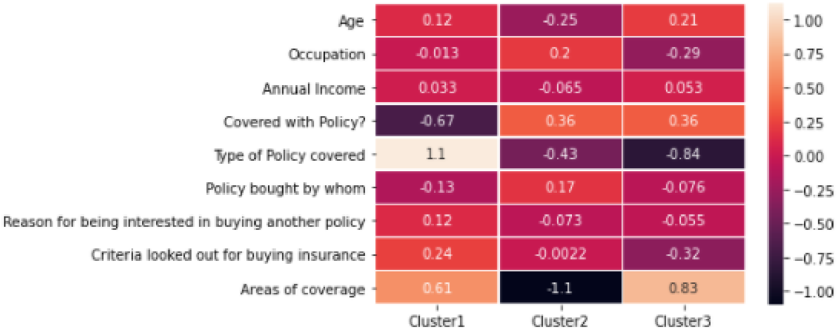


K-Means Clustering (for each dataframe)

1. Plot an ***elbow*** graph to look for elbow point (determine optimal no. of clusters)
2. Elbow is usually the point where **distortions** start to have **diminishing returns when k increases**
3. Determined k to be **3** as decrease in SSE from 3 to 4 not as large as compared to 2 to 3
4. Look for ***cluster centers*** to model the data
5. Calculate ***z-score*** for profiling to see the ranking of the variables in each cluster
6. Convert encoded values back to its original variables to derive insights

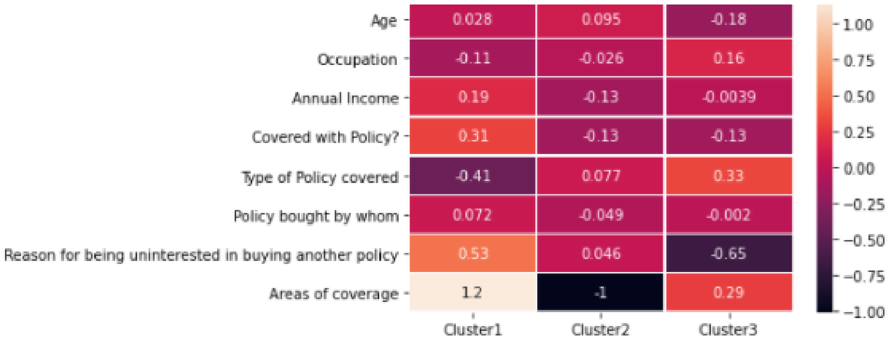
RESULTS (CLUSTERING)

Interested in buying policies



- Cluster 1: Type of policy covered
- Cluster 2: Areas of coverage
- Cluster 3: Type of policy covered

NOT interested in buying policies



- Cluster 1: Reason for being uninterested in buying another policy
- Cluster 2: Areas of coverage
- Cluster 3: Areas of coverage

RESULTS (CLUSTERING)

For people **interested** in buying policies, the top option chosen for each variable in each cluster:

	Cluster 1 (24)	Cluster 2 (27)	Cluster 3 (18)
Variable	Options	Options	Options
Type of Policy Covered	Life	Health & Life	Health
Areas of Coverage Hope to See	Sickness	Pandemic	Sickness
Policy Bought By Whom	Parents	Parents	Parents
Reason for Being Interested in Buying Another Policy	Future precautions	Future Precautions	Future Precautions
Occupation	Student	Student	Student
Criteria Looked Out For	Coverage	Coverage	Affordability
Annual Income	No income	No income	No income

RESULTS (CLUSTERING)

For people **NOT interested** in buying policies, the top option chosen for each variable in each cluster:

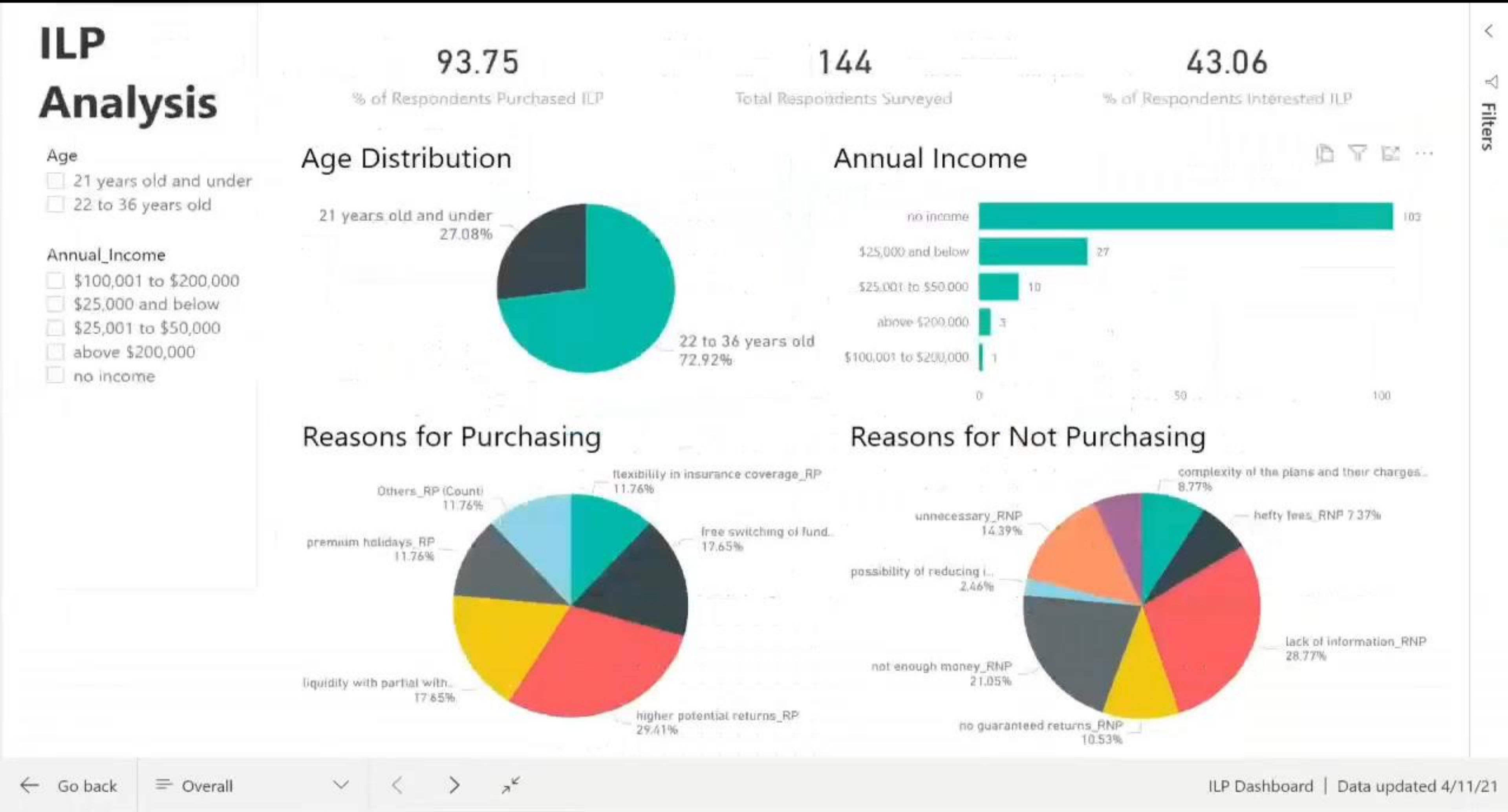
	Cluster 1 (33)	Cluster 2 (47)	Cluster 3 (30)
Variable	Options	Options	Options
Type of Policy Covered	Life	Health & Life	Life
Areas of Coverage Hope to See	Sickness	Housing/Property	Travel
Policy Bought By Whom	Parents	Parents	Parents
Reason for being uninterested in buying another policy	Family member(s) have already bought it	Family member(s) have already bought it	Family member(s) have already bought it, Unnecessary
Occupation	Student	Student	Student
Annual income	No income	No income	No income

Investments Linked Policies

NLP Insights

NLP Approach

- **Data Collection**: Google Forms
- **NLP Approach**: Steps taken are similar to what we have done previously

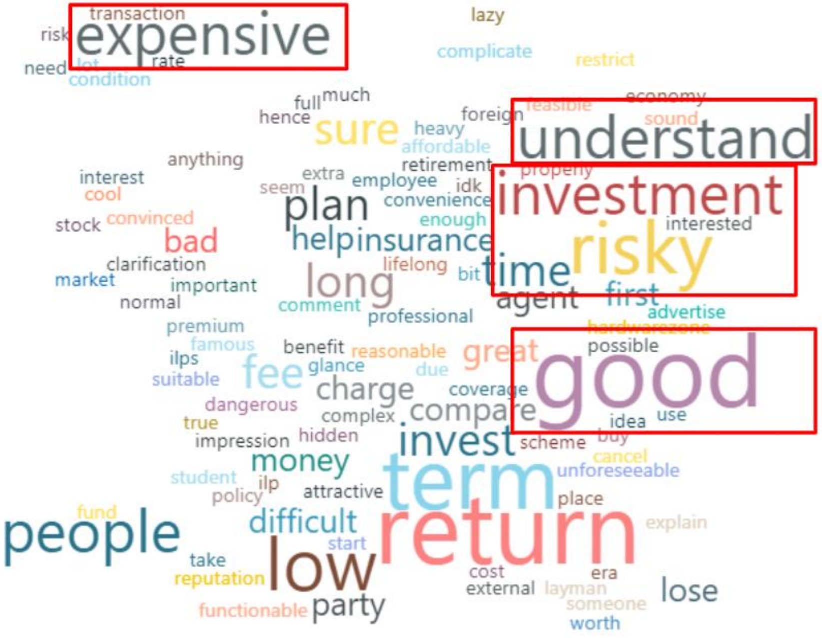

ILP
Analysis
93.75
% of Respondents Purchased ILP
144
Total Respondents Surveyed
43.06
% of Respondents Interested ILP
Age
21 years old and under
22 to 36 years old
Annual_Income
$100,001 to $200,000
$25,000 and below
$25,001 to $50,000
above $200,000
no income
Filters
Age Distribution
21 years old and under
27.08%
22 to 36 years old
72.92%
Annual Income
no income
103
$25,000 and below
27
$25,001 to $50,000
10
above $200,000
3
$100,001 to $200,000
1
0
50
100
Reasons for Purchasing
flexibility in insurance coverage_RP
11.76%
Others_RP (Count)
11.76%
free switching of fund..
17.65%
premium holidays_RP
11.76%
liquidity with partial with..
17.65%
higher potential returns_RP
29.41%
Reasons for Not Purchasing
complexity of the plans and their charges..
8.77%
unnecessary_RNP
14.39%
hefty fees_RNP 7.37%
possibility of reducing i..
2.46%
lack of information_RNP
28.77%
not enough money_RNP
21.05%
no guaranteed returns_RNP
10.53%
Go back
Overall
ILP Dashboard
Data updated 4/11/21

INSIGHTS

<u>Opinions on ILPs</u>

Qn: What do you think of ILPs in general?

- Good but risky investment
- Expensive
- Prefer a more layman's explanation
- **Recommendation**:
  - Put up informative & simple write-ups on Seedly/Social Media
  - Include benefits & possible risk



BUSINESS VALUE TO PRUDENTIAL

- Provide a **clearer understanding of Prudential's target audience** through
    - **Understanding** their **opinions on insurance, ILPs and needs** (i.e. affordability, coverage)
    - **Distinguishing** the different **customer segments** in GenY/Z
        - For Cluster 2 of customers who are interested in buying another policy
            - Pandemic→ Focus on recommending pandemic related policies
    - Identifying the different **platforms and channels to reach out** to Gen Y/Z regarding ILPs
    - **Clear** and **interactive** visualisation via **PowerBI**

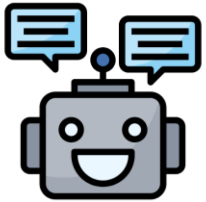
Enhancing Customers Engagement using IBM Watson Chatbot

Chatbot

RECAP: PROJECT SCOPE

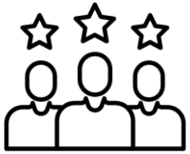
Improving The Customer (Gen Y and Z) Journey In Insurance

Customer Acquisition through Targeted Marketing



Enhancing Customers Engagement using IBM Watson Chatbot

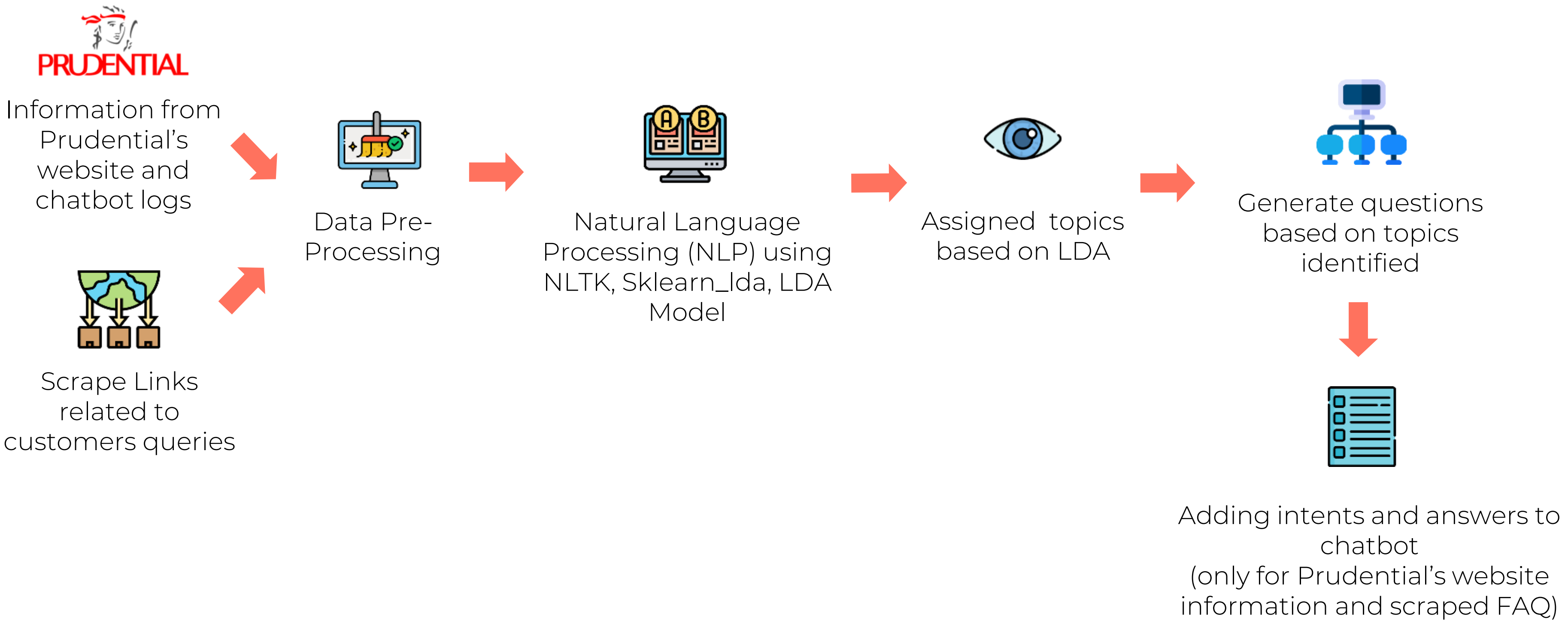
Customer Risk Assessment using Machine Learning Models



Increasing Customer Retention using Machine Learning and Clustering models

CHATBOT (PROTOTYPE)

Solution Architecture



TEXT ANALYSIS

Solution Architecture

Lemmatization using **WordNetLemmatizer** in **NLTK stem**

→

Tokenization using **RegexpTokenizer** in **NLTK**

→

Remove stop words using words in **NLTK corpus**

→

CountVectorizer using **Sklearn feature extraction text**

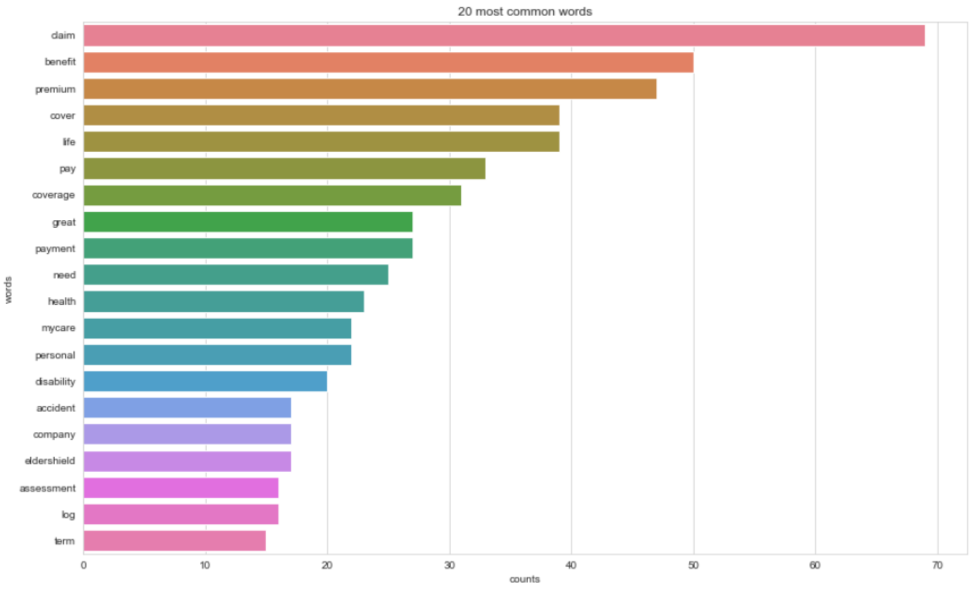
→

LDA using LDA in **sklearn decomposition**

Online scraped data & Prudential Data

NLP Results & Insights

TOPIC MODELLING RESULTS (SCRAPPED QUESTIONS)



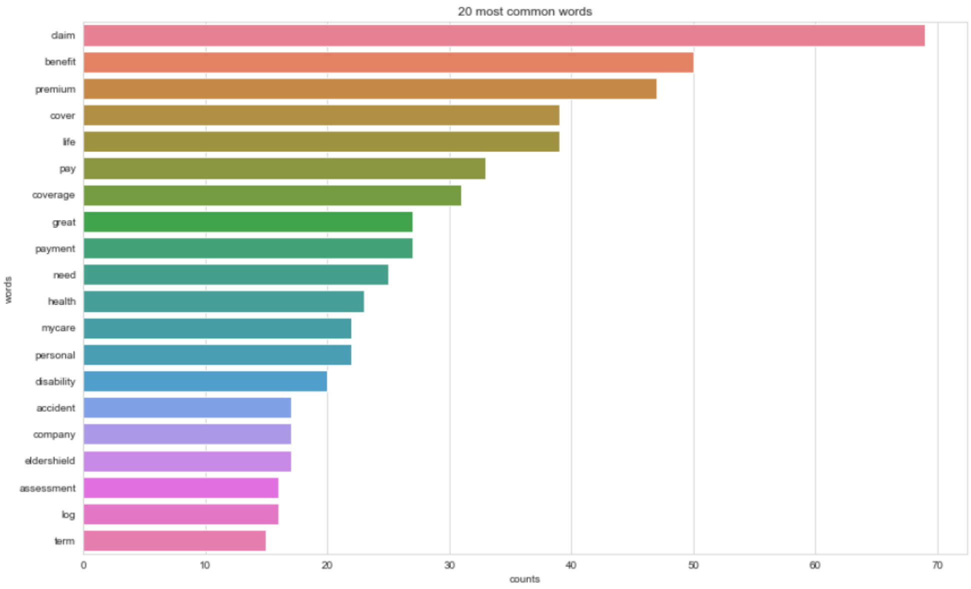

```
Topics found via LDA:

Topic #0:
need log medical health loan claim doctor eastern great giro

Topic #1:
benefit life claim great coverage company receive cash supremehealth value

Topic #2:
premium cover pay claim mycare personal eldershield accident disability payment
```

- Topic 0: Medical Insurance Coverage
- Topic 1: Life Insurance
- Topic 2: Disability Accident Insurance

Eg. of questions added to chatbot from topic modelling:

1) What is medical insurance?
2) Benefits of having a life insurance
3) What is disability insurance?

TOPIC MODELLING RESULTS (PRUDENTIAL QUESTIONS)




```
Topics found via LDA:

Topic #0:
claim form change withdrawal life accident email assure surrender medical

Topic #1:
prushield premium pruextra premier client hospital update shield pay age

Topic #2:
cover benefit table pass extra crisis charge rule illness definition
```

- Topic 0: Claims Services
- Topic 1: Medical Policies
- Topic 2: Policies Coverage

Eg. of questions that could be added to chatbot from topic modelling:

1) How to submit a claim?
2) What kind of medical policy does prudential have?
3) What is the coverage for PruPersonal Accident?

INSIGHTS

LDA models

Scraped Questions

- Topic 0: Medical Insurance Coverage
- Topic 1: Life Insurance
- Topic 2: Disability Accident Insurance

Prudential's Data

- Topic 0: Claims Services
- Topic 1: Medical Policies
- Topic 2: Policies Coverage

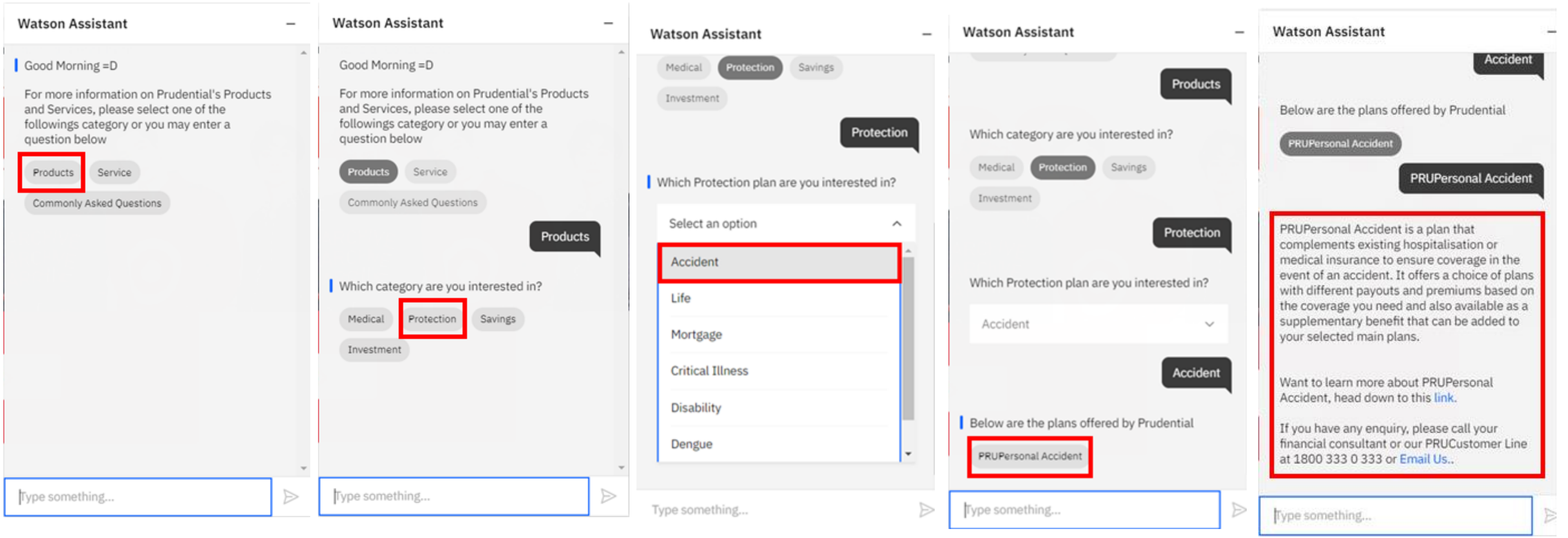
- **Area to focus on**
  - Medical Insurance
  - Life insurance
  - Disability Insurance
  - Claim Services

Chatbot (Prototype)

Use case scenarios

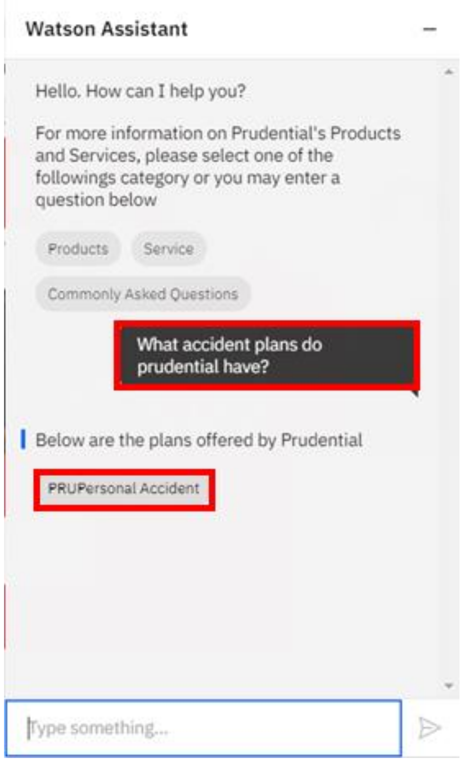
USE CASE SCENARIOS

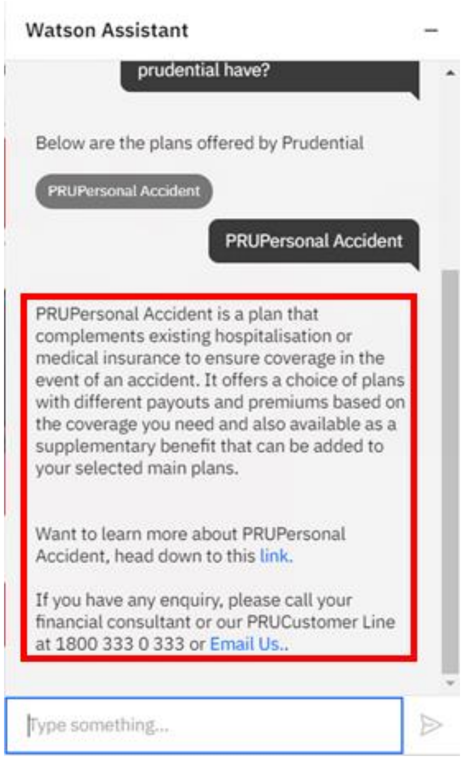
Use case #1: Prudential's potential customer wishes to know more about the different accident plans that Prudential offers. (Structured Approach)



USE CASE SCENARIOS

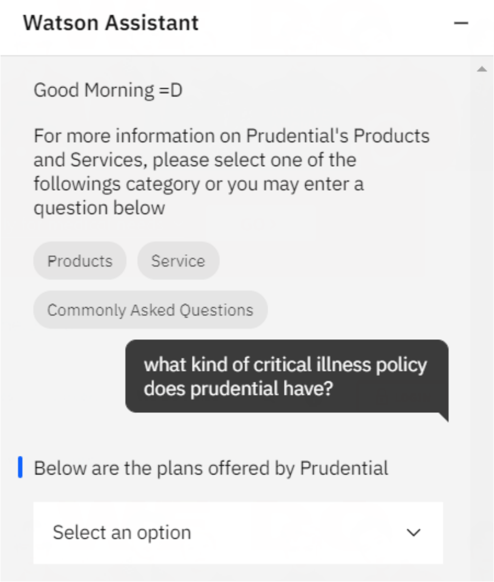
Use case #2: Prudential's potential customer wishes to know more about the different accident plans that Prudential offers. (Question Approach)

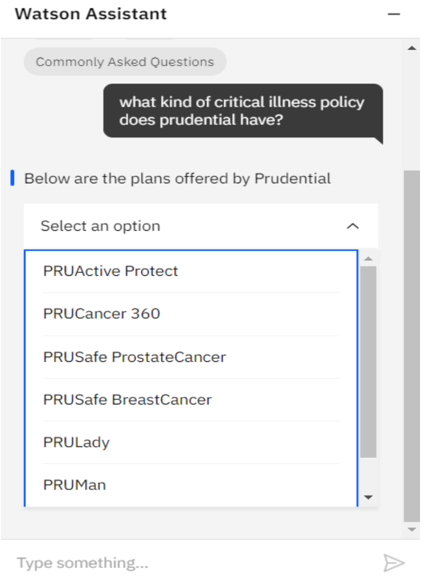




TESTING

Questions	Intents	Pass/Fail
critical illness	choose_criticalillness	Pass
What kind of critical illness policy does prudential have?	choose_criticalillness	Pass
what ci plans prudential have?	choose_criticalillness	Pass
im interested in ci	choose_criticalillness	Pass
CI	choose_criticalillness	Pass
im interested in critical illness	choose_criticalillness	Pass
What critical illness plans do prudential have?	choose_criticalillness	Pass
dengue	choose_dengue	Pass
im interested in dengue policy	choose_dengue	Pass
What dengue plans do prudential have?	choose_dengue	Pass
What kind of dengue policy does prudential have?	choose_dengue	Pass
mosquito	choose_dengue	Pass





BUSINESS VALUE TO PRUDENTIAL

Python Codes for scrapping

- **Stay updated** on the popular topics discussed amongst the general public
- Dispenses with the need for scraping from scratch

IBM Watson Chatbot (Prototype)

- Allow Prudential's customers to **efficiently obtain answers to their questions**
    - Omits using traditional means
- Provides **easy integration** with Prudential's website
- Utilize our chatbot as a **secondary reference** to **enhance their chatbot**

Customer Risk Management using Machine Learning Models

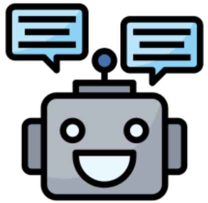
Increasing Customer Retention using Machine Learning and Clustering Models

Predictive Models

RECAP: PROJECT SCOPE

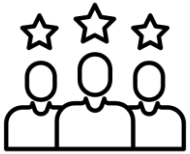
Improving The Customer (Gen Y and Z) Journey In Insurance

Customer Acquisition through Targeted Marketing



Enhancing Customers Engagement using IBM Watson Chatbot

Customer Risk Assessment using Machine Learning Models



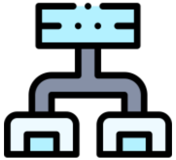
Increasing Customer Retention using Machine Learning and Clustering models

PREDICTIVE MODELS (POC)

Solution Architecture

Obtain dataset from **Kaggle**

Data Preprocessing
(Check for NA, Data standardisation, Feature selection)

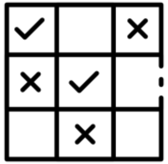


Split dataset into **Train, Validation and Test**

Balance imbalanced data with **SMOTE** (training data only)

Model fitting and selection based on **Accuracy, F-Score, Precision, Recall**

Optimise models by **tuning hyperparameters** using **GridSearchCV**



Used **KFold** to **cross validate** the generalizability of the model

DATASETS

MachineHack Insurance Churn (Customer Retention)

33,908

Observations

17

Fields

Prudential Life Insurance Assessment (Customer Risk Assessment)

59,381

Observations

127

Fields

DATASETS

MachineHack Insurance Churn (Customer Retention)

- Columns are anonymised:
  - feature_0 to feature_6 (continuous)
  - feature_7 to feature_9, feature_13 to feature_15 (categorical)
  - feature _10 to feature_12 (categorical)
  - labels (categorical) → Target Variable (0 means retained, 1 means churned)

feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	feature_11	feature_12	feature_13	feature_14	feature_15	labels
-0.27651	-0.42443	1.344997	-0.01228	0.07623	1.076648	0.182198	3	0	1	0	0	0	0	10	2	1
0.853573	0.150991	0.503892	-0.97918	-0.56935	-0.41145	-0.25194	4	1	2	0	1	0	0	0	3	0
0.947747	-0.17383	1.825628	-0.70348	0.07623	-0.41145	-0.25194	6	1	2	0	0	0	0	5	3	0
0.853573	-0.3814	0.984523	-0.03946	-0.56935	-0.41145	-0.25194	4	0	2	0	1	0	0	5	3	0
1.324443	1.590527	-1.17832	-0.09771	-0.24656	-0.41145	-0.25194	0	1	1	0	0	0	0	8	3	0
1.418617	-0.44742	1.344997	0.154691	-0.56935	0.707119	3.221163	4	1	1	0	0	0	0	10	1	1
0.288529	-0.32229	-0.81784	-0.64523	-0.56935	-0.41145	-0.25194	9	1	1	0	1	0	0	5	3	0
0.006007	-0.33214	-0.81784	-0.32293	-0.56935	-0.41145	-0.25194	9	1	1	0	1	0	0	5	3	0
-1.50078	0.02323	-0.0969	1.016744	-0.56935	-0.41145	-0.25194	8	2	2	0	1	0	0	8	3	0

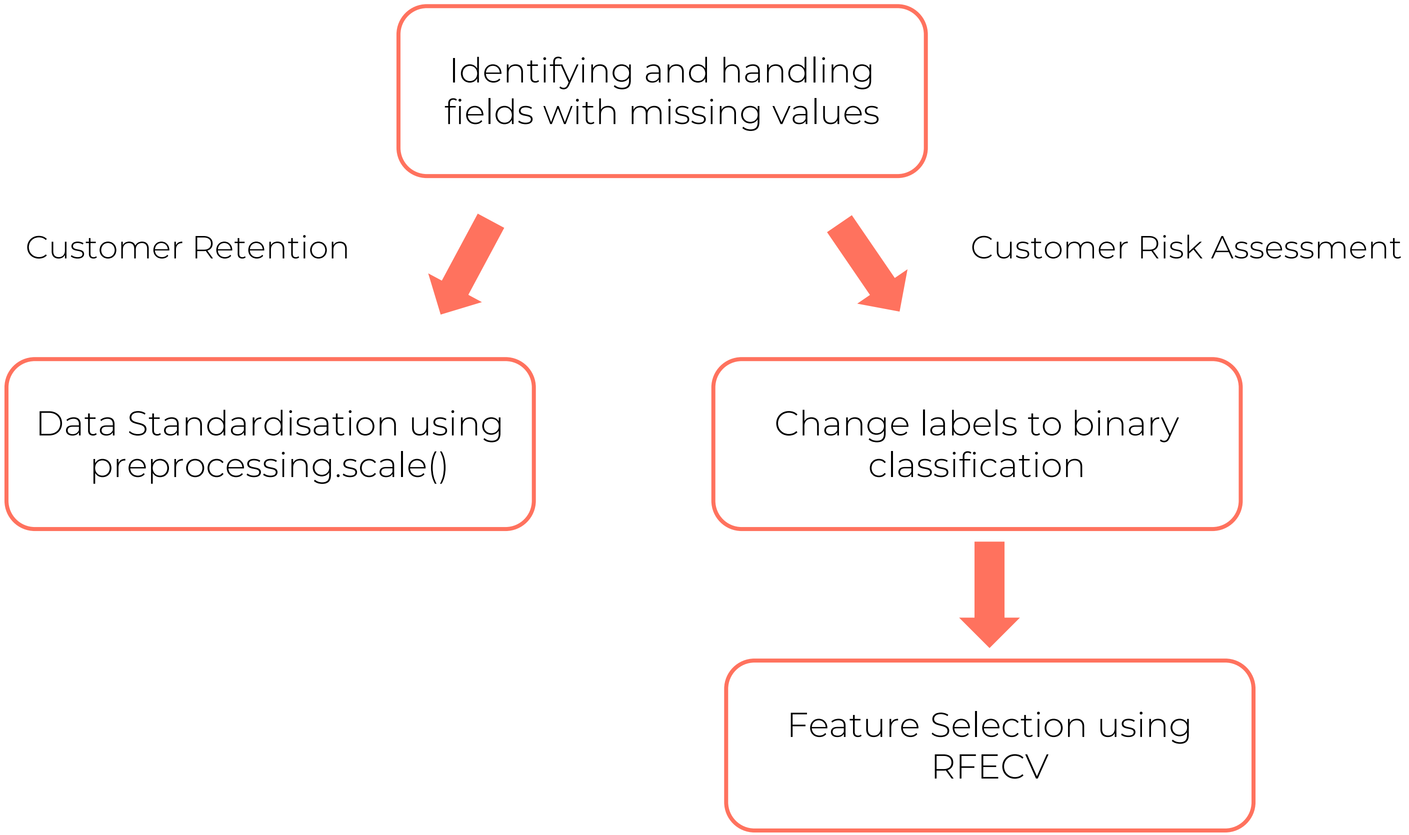
DATASETS

Prudential Life Insurance Assessment (Customer Risk Assessment)

Data Fields		
Id	BMI	Medical_History_1-41
Product_Info_1-7	Employment_Info_1-6	Medical_Keyword_1-48
Ins_Age	InsuredInfo_1-6	Response (1-8 , High risk - No Risk)
Ht	Insurance_History_1-9	
Wt	Family_Hist_1-5	

Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	Medical_H	label
3	1	3	2	3	1	3	1	2	2	1	3	3	3	8
3	1	3	2	3	3	1	1	2	2	1	3	3	1	4
3	1	3	2	3	3	3	1	3	2	1	3	3	1	8
3	1	3	2	3	3	3	1	2	2	1	3	3	1	8
3	1	3	2	3	3	3	1	3	2	1	3	3	1	8
3	1	3	2	3	3	1	1	2	2	1	3	3	3	8
3	1	1	2	3	3	3	1	2	2	1	3	3	3	8

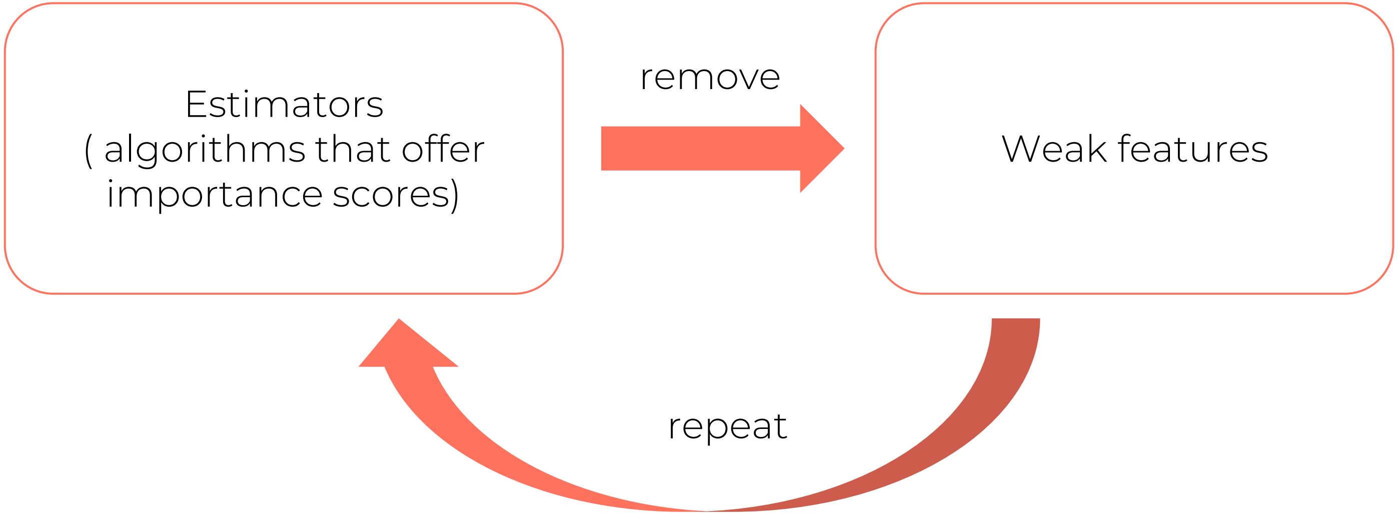
DATA PREPROCESSING



RFECV

Prudential Life Insurance Assessment (Customer Risk Assessment)

How it works?



FEATURE SELECTION

Prudential Life Insurance Assessment (Customer Risk Assessment)

RFECV- Recursive Feature Elimination and Cross-Validation Selection

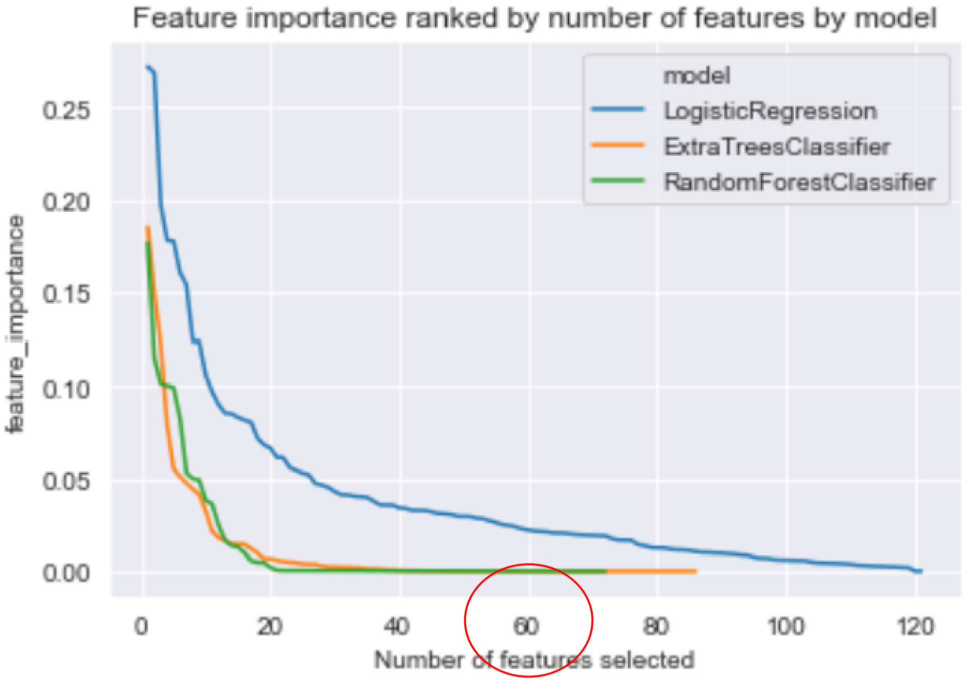
Model	Accuracy in Training set	Accuracy in Validation set
Logistic Regression	74.7%	75.1%
Extra Trees Classifier	67.1%	66.8%
Random Forest Classifier	67.1%	66.8%

Model with highest accuracy score is used as the base for feature importance ranking

FEATURE SELECTION

Prudential Life Insurance Assessment (Customer Risk Assessment)

RFECV- Recursive Feature Elimination and Cross-Validation Selection



127

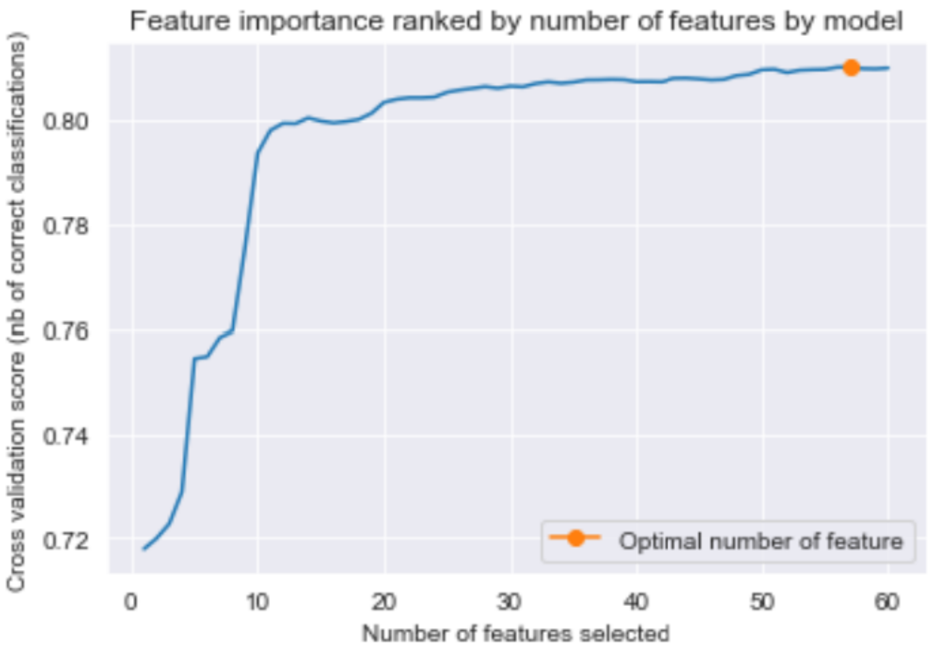
60

Fields

FEATURE SELECTION

Prudential Life Insurance Assessment (Customer Risk Assessment)

RFECV- Recursive Feature Elimination and Cross-Validation Selection



60

58

Fields

SPLIT EACH DATASET INTO TRAIN, TEST & VALIDATION

- Sklearn's train_test_split
- Train: Test = 80 : 20
- Within Train, further split Train : Validation = 80 : 20

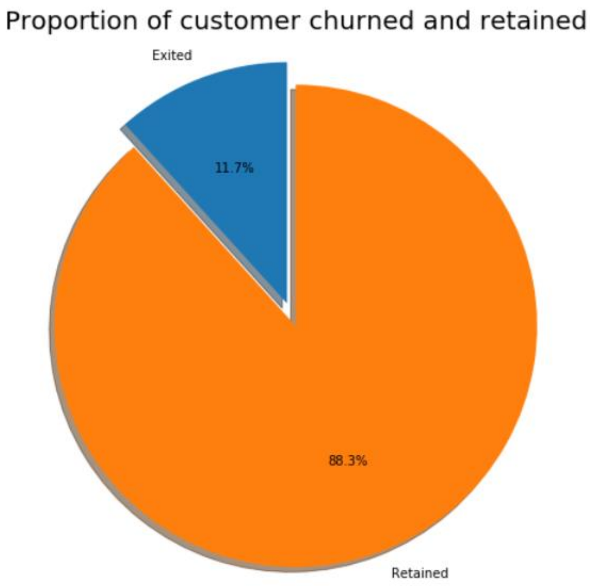
Train dataset
- To train models

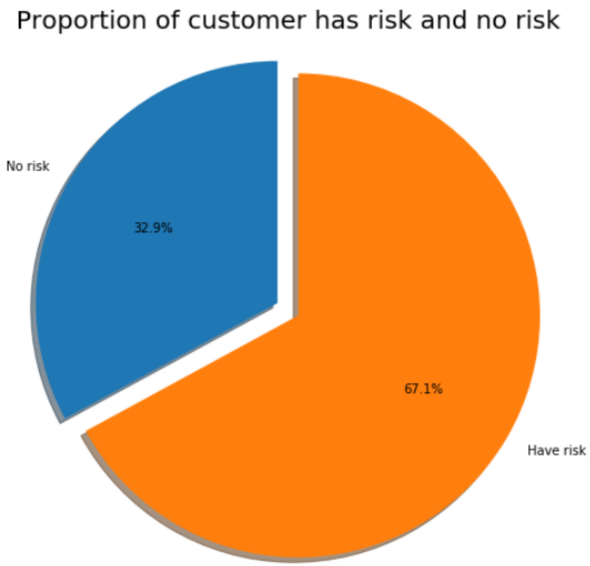
Validation dataset
- To provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters

Test dataset
- To test optimised model performance on unseen data

BALANCING LABELS WITH SMOTE

- **Synthetic Minority Oversampling Technique**
- Dataset's labels are highly **imbalanced**
  - Results in **poor performance** of machine learning models built
  - Creates a **bias** where models tends to predict the majority class





MACHINE LEARNING TECHNIQUE SELECTION

Team’s Goal

- Produce a **well-generalised model** that can work well with different datasets
- Ensure that the model built would also **work well on Prudential’s dataset**

Ensemble Methods

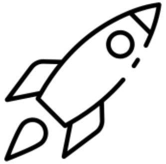
- A machine learning technique that **combines several base models** in order to produce **one optimal predictive model**
- Result in **well-generalised** models and **reduces the risk of overfitting**
    - Evaluate the features that would give better generalised results
        - Records that are wrongly classified will have their weights increased
        - Records that are correctly classified will have their weights decreased

MODEL FITTING & SELECTION

Binary Classification using Ensemble Methods

Random Forest Classifier

Extra Trees Classifier



Adaboost Classifier

XGBoost Classifier

RESULTS ON VALIDATION DATASET

MachineHack Insurance Churn (Customer Retention)

Model	Accuracy	Precision	Recall	F-Score
Random Forest	88.7%	73.1%	76.7%	74.7%
Extra Trees	88.8%	73.2%	76.1%	74.5%
Adaboost	85.6%	69.2%	78.7%	72.2%
XGBoost	86.7%	71.1%	81.6%	74.5%

RESULTS ON VALIDATION DATASET

Prudential Life Insurance Assessment (Customer Risk Assessment)

Model	Accuracy	Precision	Recall	F-Score
Random Forest	81.5%	79.1%	81.1%	79.8%
Extra Trees	81.2%	78.8%	80.4%	79.4%
Adaboost	81.8%	79.4%	80.3%	79.8%
XGBoost	81.9%	79.5%	80.8%	80.1%

MODEL OPTIMISATION ON VALIDATION DATASET

Regularisation

- Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
  - L1: alpha
  - L2: lambda

- Performed on XGBoost model only
  - Parameters: reg_alpha, reg_lambda

- Tree models prevent overfitting by controlling hyperparameters such as maximum depth of the tree and minimum size of a leaf (next section)

Model	XGBoost	
	Customer Retention Model	Customer Risk Assessment Model
Technique	L1 Regularisation	
Results	Best parameters: • Alpha: 0.70	Best parameters: • Alpha: 0.80

MODEL OPTIMISATION ON VALIDATION DATASET

Hyperparameter Tuning

Model	Customer Retention Model	Customer Risk Assessment Model
Technique	**SKlearn GridSearchCV** ● Procedure: a. Using GridSearchCV to extract different parameters b. Indicating the different variations for the model to train and understand which parameters will achieve better performance c. Re-applying the changed parameters to re-train the model **SKlearn RandomizedSearchCV** ● Procedure: a. Define a grid of hyperparameter ranges b. Randomly sample from the grid c. Perform K-Fold CV with each combination of values.	

MODEL OPTIMISATION ON VALIDATION DATASET

<table>
<tr><td colspan="2">Model</td><td>Customer Retention Model</td><td>Customer Risk Assessment Model</td></tr>
<tr><td rowspan="2">Results</td><td>XGBoost</td><td>Best parameters:
● colsample_bytree: 0.4
● learning_rate: 0.1
● max_depth: 7
● reg_alpha: 0.7 (L1 Regularisation)
Accuracy: 92.5% (+ ~6%)</td><td>Best parameters:
● colsample_bytree: 0.4
● learning_rate: 0.1
● max_depth: 7
● reg_alpha: 0.8 (L1 Regularisation)
Accuracy: 84.6% (+ ~ 2%)</td></tr>
<tr><td>Random Forest</td><td>Best parameters:
● n_estimators: 50
● criterion: “gini”
● min_samples_split: 5
● min_samples_leaf: 2
● max_features: 'sqrt'
● max_depth: None
● bootstrap: False
Accuracy: 93.3% (+ ~4%)</td><td>Best parameters:
● n_estimators= 400
● criterion="gini"
● min_samples_split= 5
● min_samples_leaf= 1
● max_features= 'sqrt'
● max_depth= None
● bootstrap= False
Accuracy: 81.6% (+ 0.1%)</td></tr>
</table>

MODEL OPTIMISATION ON VALIDATION DATASET

<table>
<tr><td colspan="2">Model</td><td>Customer Retention Model</td><td>Customer Risk Assessment Model</td></tr>
<tr><td rowspan="2">Results</td><td>Adaboost</td><td>NA</td><td>Best parameters:
● n_estimators: 225
● learning_rate: 0.3
Accuracy: 82% (+ 0.2%)</td></tr>
<tr><td>Extra Trees</td><td>Best parameters:
● criterion: ‘entropy’
● max_depth': 32
● max_features: ‘auto’
● n_estimators:100
Accuracy: 94.3% (+ ~6%)</td><td>NA</td></tr>
</table>

CROSS VALIDATION ON VALIDATION DATASET

SKlearn K-Fold Cross Validation

- A resampling procedure used to evaluate machine learning models on unseen data

	Cross Validation Accuracy	
Model	Customer Retention	Customer Risk Assessment
XGBoost	90.2%	81.6%
Random Forest	90.1%	81.5%
Extra Trees	89.9%	81.6%

FINAL RESULTS ON TEST DATASET

Customer Retention Model

Model	Parameters	Accuracy	Precision	Recall	F-Score
XGBoost	'colsample_bytree': 0.4 'learning_rate': 0.1 'max_depth': 7 'reg_alpha': 0.7 (L1 Regularisation)	89.0%	74.5%	76.4%	75.4%
Extra Trees	'criterion': 'gini' 'max_depth': 32 'max_features': 'sqrt' 'n_estimators': 50	88.6%	73.7%	76.6%	75.0%
Random Forest	n_estimators: 50 criterion = "gini" min_samples_split: 5 min_samples_leaf: 2 max_features: 'sqrt' max_depth: None bootstrap: False	88.4%	73.2%	75.8%	74.4%

FINAL RESULTS ON TEST DATASET

Customer Risk Assessment Model

Model	Parameters	Accuracy	Precision	Recall	F-Score
XGBoost	colsample_bytree: 0.4 learning_rate: 0.1 max_depth: 7 reg_alpha: 0.8 (L1 Regularisation)	81.9%	79.5%	81.9%	80.3%
AdaBoost	n_estimators=225 learning_rate =0.3	82.7%	80.1%	81.1%	80.5%
Random Forest	n_estimators= 400 criterion="gini" min_samples_split= 5 min_samples_leaf= 1 max_features= 'sqrt' max_depth= None bootstrap= False	82.2%	79.7%	81.4%	80.4%

SIMULATING REAL LIFE ENVIRONMENT

- Noise was added to the dataset to simulate datasets in real life.
- Added to the continuous features of **test dataset**

Customer Retention Model

Model	Accuracy	Precision	Recall	F-Score
XGBoost	59.3%	57.0%	66.3%	51.2%
Random Forest	70.6%	58.1%	66.5%	57.4%
Extra Trees	80.8% (9% decrease)	62.8%	69.1%	64.6%

Customer Risk Assessment Model

Model	Accuracy	Precision	Recall	F-Score
XGBoost	70.0% (12% decrease)	65.0%	59.8%	59.9%
Random Forest	69.9%	65.8%	57.2%	56.0%
AdaBoost	65.0%	56.5%	54.4%	53.8%

127.0.0.1:5000

Apps | Bookmarks | Suggested Sites | OASIS | Online Acce... | eLearn - Singapore... | IS453: G1-G2 | Slack | phpMyAdmin | Visualize Python, Ja... | Library Genesis | » | Reading list

SMU IS483 Data Ninjas x Prudential

User Guide

Customer Risk Assessment Model

Upload Training File here

Choose File No file chosen

Get Feature Importance

Darren Png, Neo Jia Ying, Nor Aisyah, Tay Yu Liang, Wong Wei Ling, Yeo Hui Xin

TESTING

- To make sure requirements are met and application has no bugs:
    - Test cases were created
    - Flask application was tested against test cases

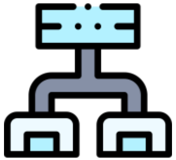
S/N	Page Name(as per requirements document)	Description	Test Inputs	Test Procedure	Expected Results	Pass/Fail
1	Home page	Validate that user with missing required file is unable to proceed	File input: No file uploaded	At home page: Click on upload without uploading file	Error will be shown	Pass
2	Home page	Validate that user with correct required file is able to train the model	File input: train.csv	At home page: -Upload train.csv -Click on upload	Upload file successful. User directed to Results/Summary Page.	Pass
3	Results/Summary	Validate that user with missing required file is unable to proceed	File input: No file uploaded	At home page: Click on upload without uploading file	Error will be shown	Pass
4	Results/Summary - Model1	Validate that user with correct required file is able to obtain results from trained model	File input: test.csv	At home page: -Upload test.csv -Click on upload	Upload file successful. Results of model 1 shown.	Pass
5	Results/Summary - Model2	Validate that user with correct required file is able to obtain results from trained model	File input: test.csv	At home page: -Upload test.csv -Click on upload	Upload file successful. Results of model 2 shown.	Pass

CLUSTERING MODEL (POC)

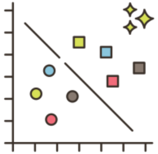
MachineHack Insurance Churn (Customer Retention)

Solution Architecture

Export dataset from best model previously and import it into Jupyter Notebook



Split dataset into **retained** and **exited**



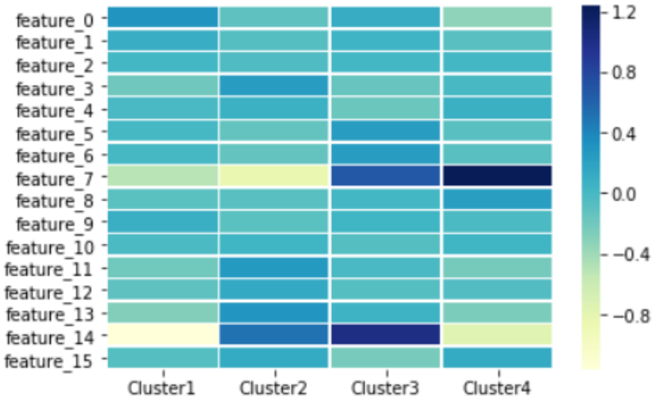
Perform **K-Means Clustering** (Elbow Graph, Assign cluster labels)

Cluster **Profiling** (Z-score)

Plot **Heatmap** to see highly ranked variables in each cluster

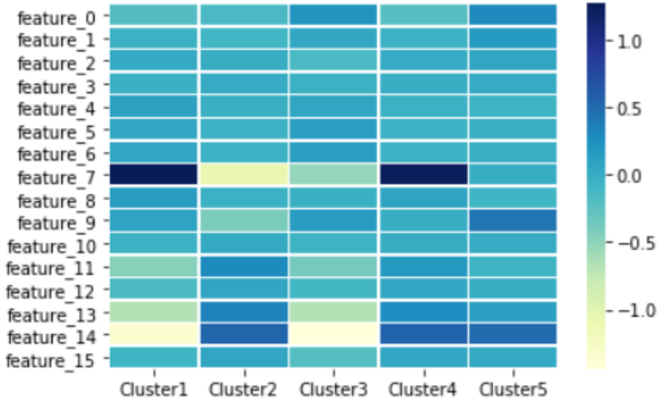
RESULTS (CLUSTERING)

For people who exited



- Cluster 1: feature_14
- Cluster 2: feature_7
- Cluster 3: feature_14 (more significant) and feature_7
- Cluster 4: feature_14 (very significant) and feature_7 (very significant)

For people who retained



- Cluster 1: feature_7 and feature_14
- Cluster 2: feature_7
- Cluster 3: feature_14
- Cluster 4: feature_7
- Cluster 5: feature_14, feature_9 and feature_0

BUSINESS VALUE TO PRUDENTIAL

Predictive Models

- **Greater proficiency** with the different data mining techniques and their performance
- POC models with **high generalisation capability**
- **Automation and Efficiency**: Flask application

Clustering Model

- Understand customers better through **identifying characteristics** of the different clusters of customers
    - **Enhance features** that affects churn rate
        - For eg. If customer service is a highly ranked variable in a cluster of customer that exited, Prudential can improve their customer service
    - Promote **complementary products** to increase sales

CHALLENGES

CHALLENGES

DATA

- Obtaining sensitive dataset from Prudential
- Anonymised column headers for the dataset from Kaggle

TECHNICAL

- Exploring new machine learning models
- Unfamiliarity with PowerBI
- Unfamiliarity with Flask

GAP ANALYSIS
&
FUTURE WORK

GAP ANALYSIS & FUTURE WORK

Predictive Models

- Trained and generalised with the aid of online datasets
    - General insights obtained
- Gap can be filled when **Prudential inputs their data into the models**, optimise the models and generate insights that are targeted at their customers

Chatbot Prototype

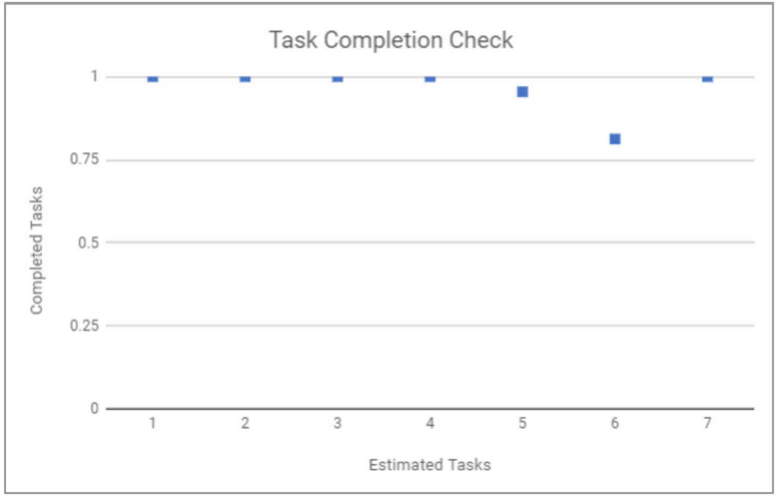
- Questions inside the chatbot are from the public and Prudential's website
    - Gap can be filled when Prudential includes **questions related to the topics** generated from their **chatbot logs** into the chatbot so that it is more tailored to their customer base

PROJECT MANAGEMENT

SCHEDULE

Iteration No	Task ID	DESCRIPTION	TYPE	PLANNED DATETIME START	PLANNED DATETIME END	DAYS SPENT PLANNED	ACTUAL DATETIME START	ACTUAL DATETIME END	DAYS SPENT ACTUAL	STATUS	LOCATION
ITERATION 1											
1	1	Meeting to discuss sourcing, scraping, formulate Google Form questions	Team Meeting - Planning, Admin, & Updates	11/1/2021	11/1/2021	1.00	11/1/2021	11/1/2021	1.00	Completed	SIS GSR 2-5
1	2	Code for Data Scraping HWZ	Programming - Coding	11/1/2021	12/1/2021	2.00	11/1/2021	22/1/2021	12.00	Completed	SIS Table
1	3	Code for Data Scraping Reddit	Programming - Coding	11/1/2021	12/1/2021	2.00	11/1/2021	11/1/2021	1.00	Completed	Home
1	4	Code for Data Scraping Seedly	Programming - Coding	11/1/2021	12/1/2021	2.00	11/1/2021	22/1/2021	12.00	Completed	Home
1	5	Data Sourcing (Problem 1, 2,3)	Data Preparation - Data Sourcing	11/1/2021	17/1/2021	7.00	11/1/2021	24/1/2021	14.00	Completed	Home
1	6	Data Scraping - Prudential Insurance	Data Preparation - Data Scraping	12/1/2021	14/1/2021	3.00	12/1/2021	22/1/2021	11.00	Completed	Home
1	39	Competitor Analysis - Life Insurance	Sentiment Analysis	20/1/21	24/1/21	5.00	22/1/2021	22/1/2021	1.00	Completed	Home
1	40	Competitor Analysis - Aviva Mindef Policy	Sentiment Analysis	20/1/21	24/1/21	5.00	23/1/2021	23/1/2021	1.00	Completed	Home
1	41	Meeting to consolidate work done at the end of the iteration	Team Meeting - Planning, Admin, & Updates	24/1/21	24/1/21	1.00	24/1/21	24/1/21	1.00	Completed	SIS GSR 2-5
ITERATION 2											
2	1	Meeting to update and set goals for this iteration	Team Meeting - Planning, Admin, & Updates	25/1/21	25/1/21	1.00	25/1/21	25/1/21	1.00	Completed	IS Lounge
2	2	Data sourcing for Problem 1 (Health, Wealth, Aspirations)	Data Preparation - Data Sourcing	25/1/21	27/1/21	3.00	25/1/21	25/1/21	1.00	Started	Home
2	44	Do up Google form for Gen Y/Z aspirations	Other	3/2/21	7/2/21	5.00	4/2/21	4/2/21	1.00	Completed	Home
2	45	Meeting to consolidate work done at the end of the iteration	Team Meeting - Knowledge Sharing	7/2/21	7/2/21	1.00	7/2/21	7/2/21	1.00	Completed	Discord
2	46	Data preparation & Model Development	Milestone - Preparation	7/2/21	7/2/21	1.00	7/2/21	7/2/21	1.00	Completed	Discord
ITERATION 3											
3	1	Meeting to update and set goals for this iteration	Team Meeting - Planning, Admin, & Updates	8/2/21	8/2/21	1.00	8/2/21	8/2/21	1.00	Completed	Discord
3	2	Meeting with Prudential for updates	Team Meeting - Planning, Admin, & Updates	8/2/21	8/2/21	1.00	8/2/21	8/2/21	1.00	Completed	MS Teams
3	18	Meeting with Prof	Team Meeting - Planning, Admin, & Updates	18/2/21	18/2/21	1.00	18/2/21	18/2/21	1.00	Completed	SIS L4
3	19	Start on Midterm slides	Milestone - Preparation	17/2/21	20/2/21	4.00	17/2/21	20/2/21	4.00	Incomplete	SIS GSR 2-7
3	20	Meeting to conclude iteration	Team Meeting - Planning, Admin, & Updates	21/2/21	21/2/21	1.00	20/2/21	20/2/21	1.00	Not started	Discord
ITERATION 4 (Preparation for Midterm Review)											
4	1	Meeting to update and set goals for this iteration	Team Meeting - Planning, Admin, & Updates	22/2/21	22/2/21	1.00			1.00		
4	2	Prepare midterm slides	Milestone - Preparation	23/2/21	2/3/21	8.00			1.00		

Planned Vs Actual



INTERNAL **MEETINGS**

- Every Sunday and Wednesday (Physically/Virtually)
- Meeting agendas include consolidating tasks, project updates, clarifications, discussions, consultation, setting tasks and goals
- Informal discussion via Telegram

To do by Sunday:
Darren: Insights (Competitor analysis + General+general Aspirations)
YL: google form insights and google form dashboard
Huixin & JY: Prob 5 risk model
WL & Aisyah: Problem 1 new data from pru if they give CSV, chatbot (scrape common qns) edited 6:31 PM

DISCORD

4 Feb meeting

BY MONDAY
• **Workout the change of scope with Pru by Monday and send after we finalised**
— Remember to write email to Prof on Monday
—Next meeting we should already be fixed on our deliverables
• Prof thinks it's okay to do POC but concerned regarding sponsor whether they are satisfied with it

PROPOSAL & PRESENTATION RELATED
• Our deliverable for midterm should be **model selection** instead of concrete model (since not on their dataset)
• Criterias that we used should be clearly stated in proposal
• The accuracy might only be for the current dataset. The model needs to be generalised, meaning any similar datasets should give around the same accuracy. —> how?
—By providing some mechanisms like transfer learning techniques apart from just providing the POC to show evidence that it can work well on another dataset
—Prof is concerned about us just comparing different models see which model is best because ultimately, what matters is the mode doing well on their dataset
• **To research on:**
—**Privacy preserving** machine learning technique to control the leakage (some leakage still)
—**secure machine learning** - you purely do not leak any data at all
—**differential privacy system**
edited 2:25 PM

MEETING MINUTES

13 Jan 2021 Wednesday 3.30pm SIS GSR 2.5

- Data Sourcing, Pre-processing and Cleaning
- Made changes to google forms
  - Craft questions
- Send out google form
- Discuss about the change of scope
- Perform more scraping (HWZ + KS)

18 Jan 2021 Monday 9pm Online Discord

- Update project scope
- Update group about progress on data cleaning & EDA (word cloud & topic modelling)
- Share and compile difficulties faced to update Jamie
- Prepare slides to present to Prudential about changes

20 Jan 2021 Wednesday 11am Online MS Teams

- Present project scope slides to Jamie (Handover)
- Sharing on difficulties
- Gather feedback/opinions from Jamie
- Update scope (TBC)

20 Jan 2021 Wednesday 330pm SIS GSR 2.5

- Assign tasks to be done before next meeting

21 Jan 2021 Thursday 2pm Prof's Office

- Update prof on the change of scope
- Update prof on progress
- Confirm project deliverables (schedule etc.)

UPDATES WITH PRUDENTIAL

[Confirmed] Prudential X SMU - Project Updates

Accept | Tentative | Decline

LJ Luis JH Aiw <luis.jh.aiw@prudential.com.s

Required TAY Yu Liang; Madhan Seduraman; Zhang Siqi; Magdalene MP Loh

Thu 15-Apr-21 3:26 PM

Optional YEO Hui Xin; Nor Aisyah Binte AJIT; WONG Wei Ling; Darren PNG Wei Xuan; NEO Jia Ying

We couldn't find this meeting in the calendar. It may have been moved or deleted.

Friday, April 16, 2021 5:30 PM-6:15 PM, ~~*(Tuesday, April 20, 2021 11:00 AM-11:45 AM)*~~ Microsoft Teams Meeting

HI All,

Setting up a mid-point check on the project updates by SMU.

Agenda

1. Aligning understanding critical items required to feed into SMU's model for useful outcomes.
2. Updates from SMU's AL/ML model
3. Updates on Prudential's dataset
4. Discussion on continuity of SMU's AL/ML

Prudential X SMU - Project Updates

Reply | Reply All | Forward

TY TAY Yu Liang

To Alice Yu

Mon 12-Apr-21 5:49 PM

Cc Luis JH Aiw; Nor Aisyah Binte AJIT; WONG Wei Ling; Darren PNG Wei Xuan; YEO Hui Xin; NEO Jia Ying

Hi Alice,

Our group wanted to update you on our progress of Gen Y/Z perspective on ILP. We had collected information from Gen Y/Z mainly through surveys and currently, we are consolidating the results and dashboards.

Thank you.

Best Regards,
Tay Yu Liang

HANDOVER

- Industry Standard
- Smooth transition
- Trackable

<u>Deliverables</u>

- Flask application for predictive models
- Chatbot prototype
- Dashboards
- User guides

<u>Source Codes</u>

- Clustering model codes
- Text Analysis/NLP codes (Topic Modelling, Lemmatization, Tokenization, POS Tagging etc)
- Predictive Model codes (Optimisation, Cross Validation etc)
- Chatbot prototype (Scrapping codes, Links scrapped)
- User guides

LEARNING OUTCOMES

- Experienced the **difference between school and working environment**
  - Uncertainties
  - Scope changes
  - Tight timeline
- Gained experience for **constructing and implementing a solution** for a large corporation's (Prudential) business operations
- **Self-learning** and **research** on aspects which were not covered in the syllabus provided in university which polished our technical knowledge
  - Going beyond our current skill set
  - **Enhancing current skills** (eg. communication, teamwork, technical skills) to adequately deal with the problems and goals within our project

THANK YOU