



HEALTH INSURANCE UPSELL USING MACHINE LEARNING

FINAL PROJECT TECHNICAL REPORT

UNIVERSITY OF NEW HAVEN

SUBMITTED BY: (GROUP 9)

1. Raghava Akula
2. Gowtham Reddy
3. Hemanth
4. Surya Manjri

SUBMITTED FOR DSCI-6002-04(FALL 2023)

DR. ARDIANA SULA



TABLE OF CONTENTS

ABSTRACT.....	3
LIST OF ABBREVIATION.....	4
INTRODUCTION.....	5
1.1 GENERAL.....	5
1.2 PURPOSE.....	6
LITERATURE SURVEY.....	8
METHODOLOGY.....	10
IMPLEMENTATION.....	13
DATA INSIGHTS.....	19
APPLICATION DEPLOYMENT.....	22
CONCLUSION.....	23
REFERENCES.....	24



Abstract

Health Insurance Businesses are in fierce competition with one another to attract new customers and retain existing ones as a result of the rise of numerous competitors and entrepreneurs.

Because of the aforementioned, no matter the size of the company, providing great customer service becomes necessary. Additionally, any company that can comprehend the wants of each of its clients would be better able to provide tailored customer care programmes and focused customer services. Structured customer service makes this comprehension possible. Customers in each sector have comparable market characteristics. In lieu of traditional market analytics, which frequently fail, particularly when the customer base is too big, big data ideas and machine learning have encouraged a greater acceptance of the automated customer segmentation approach. For this, the k-Means clustering algorithm is employed in this work. The programme is trained using a two-factor dataset of 100 patterns that were acquired from the retail industry. The sk-learn library was created for the k-Means algorithm. Characteristics of the typical number of monthly customer visits and average number of purchases.



LIST OF ABBREVIATION

ACRONYM	EXPANSION
SOM	SORTING MAP
IT	INFORMATION TECHNOLOGY
ECM	ELBOW CRITERION METHOD
SSE	SUM OF SQUARED ERRORS
AVG	AVERAGE
DF	DATA FRAME



INTRODUCTION

1.1 GENERAL

Due to the growing rivalry amongst companies and the accessibility of vast amounts of historical data, data mining techniques have become widely used in an effort to unearth critical and strategic information that is concealed in organizational data. Extracting logical information from a dataset and presenting it in a form that is understandable by humans for the purpose of decision support is known as data mining. Data systems, artificial intelligence, machine learning, and statistics are among the fields that data mining techniques differentiate. Applications for data mining encompass a wide range of fields, such as bioinformatics, financial analysis, fraud detection, weather forecasting, and customer segmentation. The use of data mining to identify customer segments in the business is crucial to this paper. The product team at Insurance All, a company that offers auto insurance to clients, is examining the viability of introducing policyholders to health insurance as a new offering. Similar to auto insurance, users of this new health insurance plan must pay a yearly premium to Insurance All in order to receive an insured amount from the company. This amount is meant to cover potential accident or vehicle damage costs. Last year, Insurance All polled roughly 380,000 clients to find out if they would be interested in signing up for a new health insurance plan. Every customer indicated whether or not they would like to purchase health insurance, and their answers, along with other customer characteristics, were stored in a database. 127,000 new customers who did not reply to the survey



were chosen by the product team to take part in a campaign wherein they will be offered the new health insurance product. The sales team will call potential customers with the offer. Nonetheless, during the campaign period, the sales team can place up to 20,000 calls.

1.2 PURPOSE

We will make one of the most significant uses of machine learning in this Data Science Project series: UPSELLING in a business. We will use Python client components for this project. The best place to look for your ideal customer is the customer division. You will receive a background on customer segmentation for this machine learning project. Subsequently, we will assess the information using which to construct the classification model. We will also see a descriptive analysis of our data and use multiple K-means algorithm types in this data science project. Proceed with the customer science project in its entirety using a Python-based algorithm learning machine. A critical application of unreadable learning is customer classification before upselling. Businesses can determine customer segments through merger strategies, which enables them to determine user bases. K-methods integration, a crucial algorithm for integrating an unlisted dataset, will be used for this machine learning project.



LITERATURE SURVEY

The business world has grown more competitive over time as a result of the need for companies like these to satisfy the requirements and desires of their current clientele in order to draw in new business and grow their clientele. It is a very challenging task for the business to determine and satisfy each customer's needs and requirements. This is due to the fact that different customers may have different needs, wants, shapes, tastes, and other characteristics. As it stands, treating every customer equally in business is a bad practice. Due to this difficulty, the idea of customer or market segmentation has gained traction. In this approach, consumers are categorized into smaller groups or segments based on shared characteristics or behaviors. Thus, breaking the market up into native groups is the process of customer segmentation.

A. BIG DATA

Big Data research has accelerated recently. The term "big data" refers to a vast amount of formal and informal data that cannot be analyzed with conventional techniques and algorithms.

Businesses store billions of data about their clients, partners, and business processes.

Additionally, millions of internally connected sensors are sent out into the real world on gadgets like smartphones and cars, gathering, producing, and transmitting data. the capacity to enhance forecasting, reduce costs, boost productivity, and enhance decision-making in a variety of domains, including finance, business transactions, national security, healthcare, education, traffic management, weather forecasting, and disaster prevention. Volume, Variability, and Speed are the three Vs that are most commonly associated with big data. Authenticity and value are additional 2Vs that make it a 5V.



B. DATA COLLECTION

The process of gathering and comparing data to specific variations in a working system allows one to assess outcomes and respond to pertinent inquiries. In all academic disciplines, including the business, humanities, and social and physical sciences, data collection is an essential component of research. All data collection aims to provide high-quality evidence so that analysis can produce answers to the submitted questions that are both deceptive and convincing. We gathered information from KAGGEL.

B. CLUSTERING DATA

The process of grouping the data in a dataset according to certain similarities is called clustering. Depending on the circumstances, a variety of algorithms can be selected and used on a dataset. It is crucial to choose the right clustering techniques because there isn't a single, all-encompassing clustering algorithm. Using the Python sklearn library, we have implemented three clustering algorithms in this project.

C. K-MEAN

The notation K-designates one of the most widely used classification algorithms.

Each data point is positioned in one of the overlapping K clusters pre-programmed into the clustering algorithm, which is based on the centroid. The clusters are formed in accordance with the hidden pattern in the data that offers the details required to assist in determining the execution process. K-means assembly can be made in a variety of ways; here, we'll employ the elbow method.



METHODOLOGY

A data science project will be developed to address this issue. A machine learning model used in this project will be able to forecast which clients might be considering purchasing health insurance. It will assist the team in choosing the most qualified clients who might be considering health insurance.

A. DATA DESCRIPTION

The data will be gathered and examined in this section. Threats or removals will be applied to the absent values. In order to understand the data, a preliminary data description will be completed. As a result, several descriptive statistics calculations will be performed, including those for kurtosis, skewness, media, fashion, median, and standard deviation.

B. FEATURE ENGINEERING

A mind map will be made in this section to help with the formulation of the hypothesis and the development of new features. These presumptions may raise the model scores and aid in exploratory data analysis.



C. DATA FILTERING

This step involves eliminating rows or columns that are not related to the business. Columns containing customer IDs, hashes, or rows containing ages that don't correspond to human ages are a few examples.

D. EXPLORATORY DATA ANALYSIS

To help with understanding the database, the section on exploratory data analysis includes univariate, bivariate, and multivariate analyses. The bivariate analysis will test the hypothesis that was developed in step 02 of the process.

E. DATA PREPARATION

The data will be ready for machine learning modeling in this fifth section. They may therefore be encoded, oversampled, subsampled, or rescaled in order to enhance the machine learning model's learning process.

F. FEATURE SELECTION

Following the data preparation in this section, Boruta and other algorithms will choose the best columns to be used in the machine learning model's training. As a result, there is less likelihood of overfitting and the database has fewer dimensions.



G. MACHINE LEARNING MODEL

The purpose of this step is to train the algorithms for machine learning and increase their capacity for data prediction. To determine the model's learning capacity, it is trained, validated, and then subjected to cross-validation.

H. FINE-TUNING THE HYPARAMETERS

After choosing the best model to use for the project, it's critical to adjust the parameters to raise the model's scores. The model performance techniques employed in step 07 remain the same.

I. CONCLUSION

In this final step, the generation capacity model is evaluated with hypothetical data. Furthermore, a few business-related queries are addressed to demonstrate the model's suitability in a business setting.

J. MODEL DEPLOY

This is the data science project's last step. Thus, the model and functions are saved to be implemented in the Flask API, and the Flask API is created in this step.



TECHNICAL INFORMATION

With the help of Python 3.x and a few Python packages for data processing, analysis, visualization, and editing, the code below was written according to the Jupyter manual.

The following code uses open source data from KAGGLE.



IMPLEMENTATION

Importing Libraries and Loading Data

```
import numpy          as np
import pandas         as pd
import seaborn       as sns
#import psycopg2      as pg
import scikitplot    as skplt
import category_encoders as ce

import sklearn.metrics as mtr
import matplotlib.pyplot as plt

from scipy          import stats
```

```
df1=pd.read_csv('C:/Users/Raghavendra Reddy/Downloads/train.csv')
```

```
df1.head()
```

	id	gender	age	driving_license	region_code	previously_insured	vehicle_age	vehicle_damage	annual_premium	policy_sales_channel	vintage	response
0	1	Male	44	1	28	0	> 2 Years	Yes	40454	26	217	1
1	2	Male	76	1	3	0	1-2 Year	No	33536	26	183	0
2	3	Male	47	1	28	0	> 2 Years	Yes	38294	26	27	1
3	4	Male	21	1	11	1	< 1 Year	No	28619	152	203	0
4	5	Female	29	1	41	1	< 1 Year	No	27496	152	39	0

```
df1.tail()
```

	id	gender	age	driving_license	region_code	previously_insured	vehicle_age	vehicle_damage	annual_premium	policy_sales_channel	vintage	response
381104	381105	Male	74	1	26	1	1-2 Year	No	30170	26	88	
381105	381106	Male	30	1	37	1	< 1 Year	No	40016	152	131	
381106	381107	Male	21	1	30	1	< 1 Year	No	35118	160	161	
381107	381108	Female	68	1	14	0	> 2 Years	Yes	44617	124	74	
381108	381109	Male	46	1	29	0	1-2 Year	No	41777	26	237	



Data Cleaning and Feature Engineering

```
df1.isna().mean()
```

```
gender          0.0
age             0.0
driving_license 0.0
region_code     0.0
previously_insured 0.0
vehicle_age     0.0
vehicle_damage  0.0
annual_premium  0.0
policy_sales_channel 0.0
vintage         0.0
response        0.0
dtype: float64
```

```
df1['previously_insured'] = df1['previously_insured'].map({0: 'no', 1: 'yes'})
df1['response'] = df1['response'].map({0: 'no', 1: 'yes'})
df1['driving_license'] = df1['driving_license'].map({0: 'no', 1: 'yes'})
df1['vehicle_damage'] = df1['vehicle_damage'].map({'Yes': 'yes', 'No': 'no'})
```

```
df1.head()
```

	gender	age	driving_license	region_code	previously_insured	vehicle_age	vehicle_damage	annual_premium	policy_sales_channel	vintage	response
0	Male	44	yes	28	no	> 2 Years	yes	40454	26	217	yes
1	Male	76	yes	3	no	1-2 Year	no	33536	26	183	no
2	Male	47	yes	28	no	> 2 Years	yes	38294	26	27	yes
3	Male	21	yes	11	yes	< 1 Year	no	28619	152	203	no
4	Female	29	yes	41	yes	< 1 Year	no	27496	152	39	no



```
cat_attributes.describe().T
```

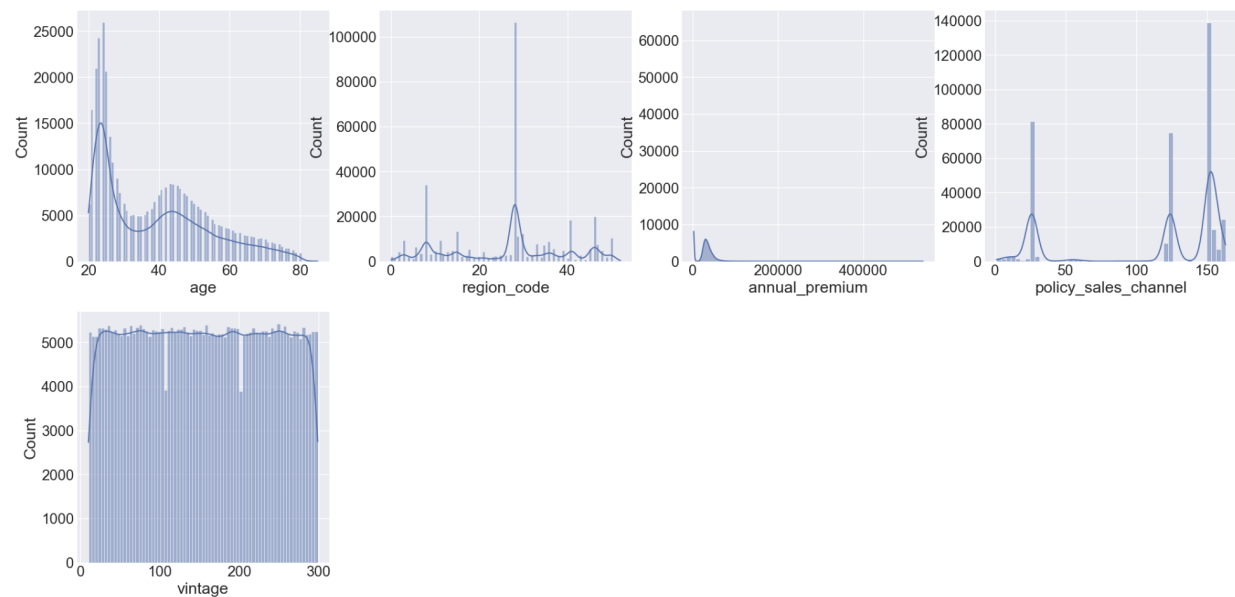
	count	unique	top	freq
gender	381109	2	Male	206089
driving_license	381109	2	yes	380297
previously_insured	381109	2	no	206481
vehicle_age	381109	3	1-2 Year	200316
vehicle_damage	381109	2	yes	192413
response	381109	2	no	334399

Univariate Analysis

```
aux1 = df4.select_dtypes(exclude='object')
columns = aux1.columns.tolist()
j = 1

for column in columns:
    plt.subplot(2, 4, j)
    sns.histplot(aux1[column], kde=True);

    j += 1
```

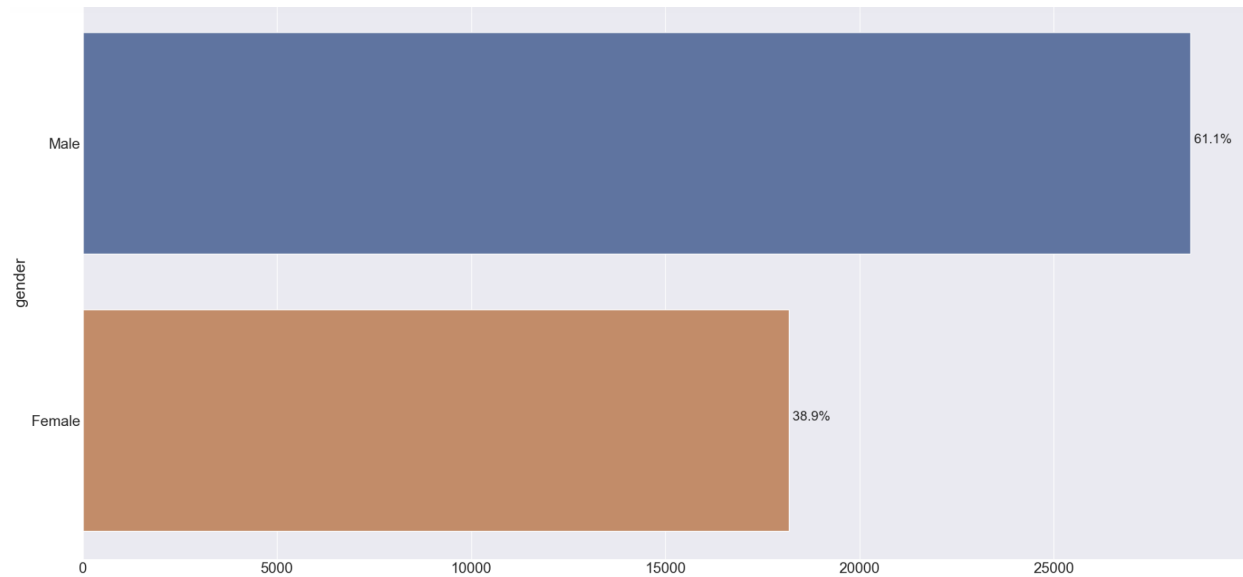




Bivariate Analysis

```
aux1 = df4[df4['response'] == 'yes']
ax = sns.countplot(y='gender', data=aux1)

total = aux1['gender'].size
for p in ax.patches:
    percentage = '{:.1f}%'.format(100 * p.get_width()/total)
    x = p.get_x() + p.get_width() + 0.02
    y = p.get_y() + p.get_height()/2
    ax.annotate(percent, (x, y))
```



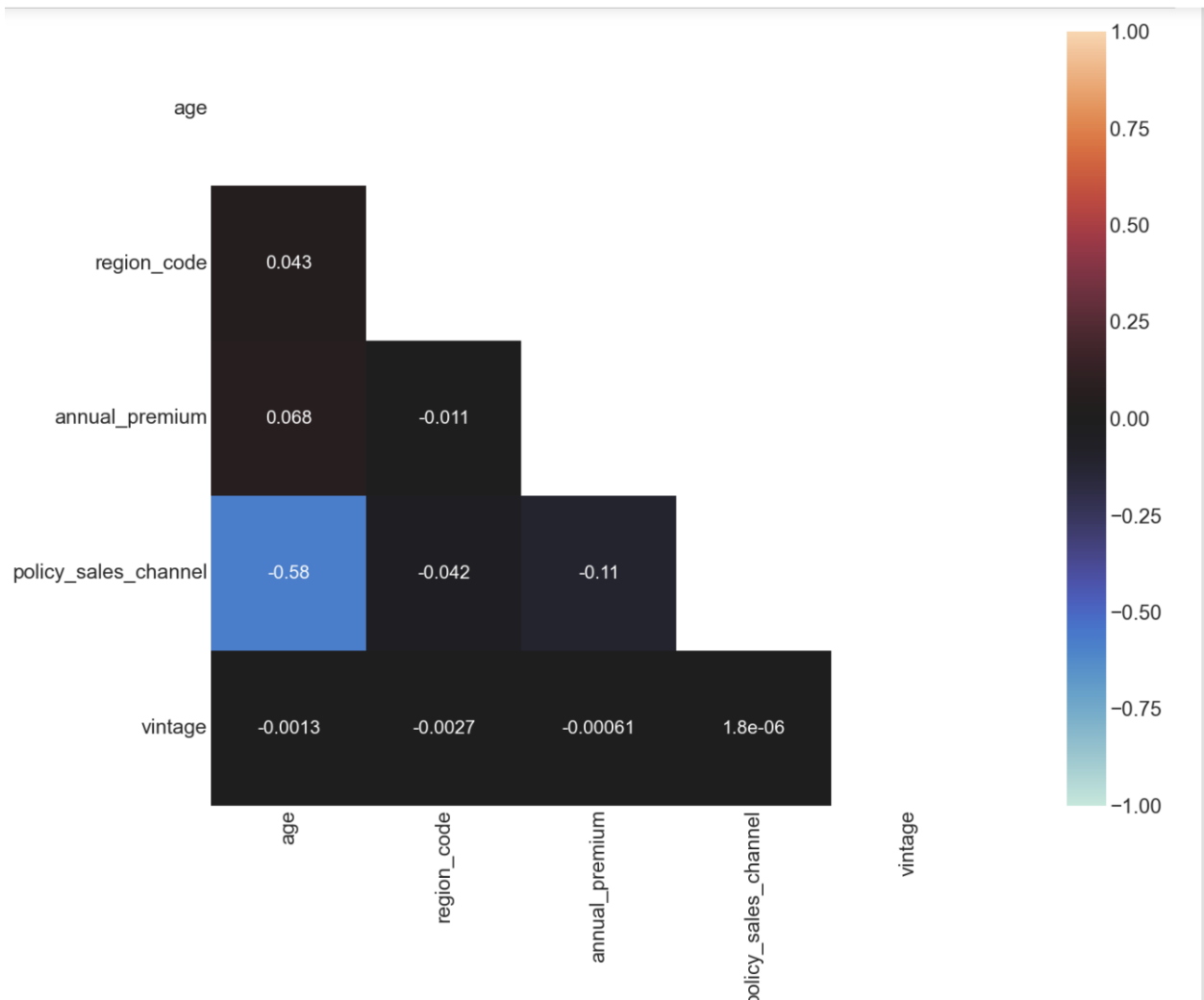


Multivariate Analysis

```
corr = aux1 = df4.select_dtypes(exclude='object').corr()

mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True

with sns.axes_style("white"):
    ax = sns.heatmap(corr, annot=True, mask=mask, vmin=-1, center=0, vmax=1, square=True)
```





Machine Learning Modeling

```
df7_params = X_params_cs
df7_test = X_test_cs
```

```
df7_train = X_train_cs
df7_valid = X_valid_cs
```

```
df7_temp = X_temp_cs
```

```
X_params, y_params = df7_params.iloc[:, :-1], df7_params.iloc[:, -1]
X_test, y_test = df7_test.iloc[:, :-1], df7_test.iloc[:, -1]
```

```
X_train, y_train = df7_train.iloc[:, :-1], df7_train.iloc[:, -1]
X_valid, y_valid = df7_valid.iloc[:, :-1], df7_valid.iloc[:, -1]
```

```
X_temp, y_temp = df7_temp.iloc[:, :-1], df7_temp.iloc[:, -1]
```

```
lg = LogisticRegression(class_weight='balanced')
lg.fit(X_train, y_train)
```

```
y_prob = lg.predict_proba(X_valid)
```

```
y_prob
```

```
array([[0.99228225, 0.00771775],
       [0.67048687, 0.32951313],
       [0.27454706, 0.72545294],
       ...,
       [0.53258921, 0.46741079],
       [0.99669209, 0.00330791],
       [0.3321847 , 0.6678153 ]])
```

```
k = int(y_valid.shape[0] / 2)
```

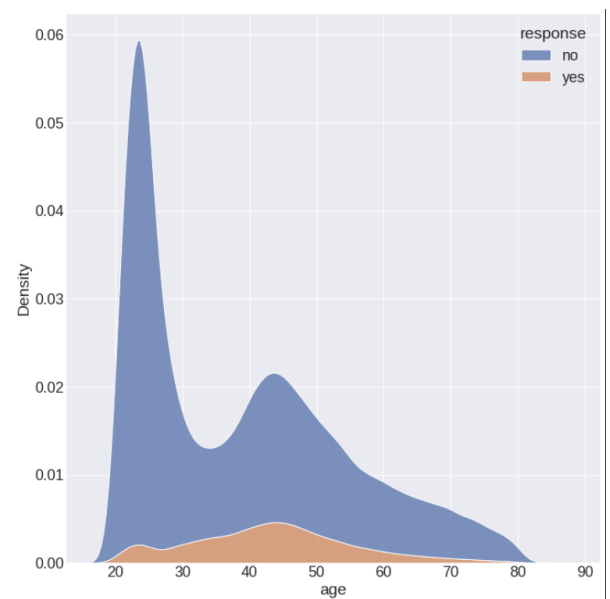
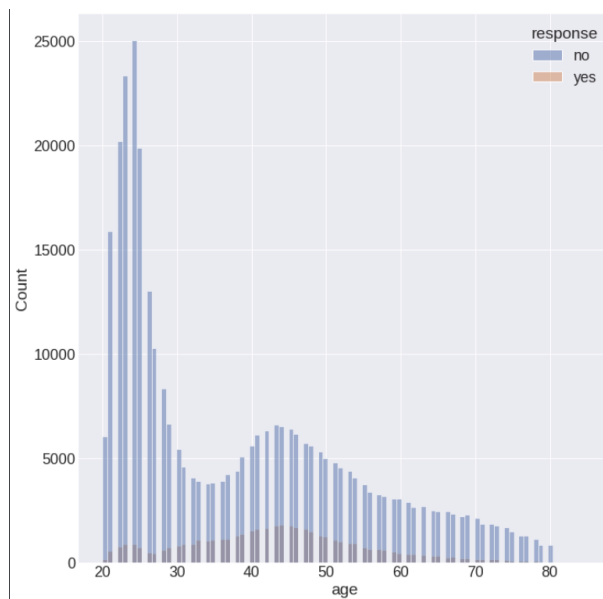
```
lg_results = ml_top_k_scores('Logistic Regression', y_valid, y_prob, k=k)
lg_results
```

	Precision_at_k	Recall_at_k	F1_at_k
Logistic Regression	0.2424	0.989	0.3894



DATA INSIGHTS

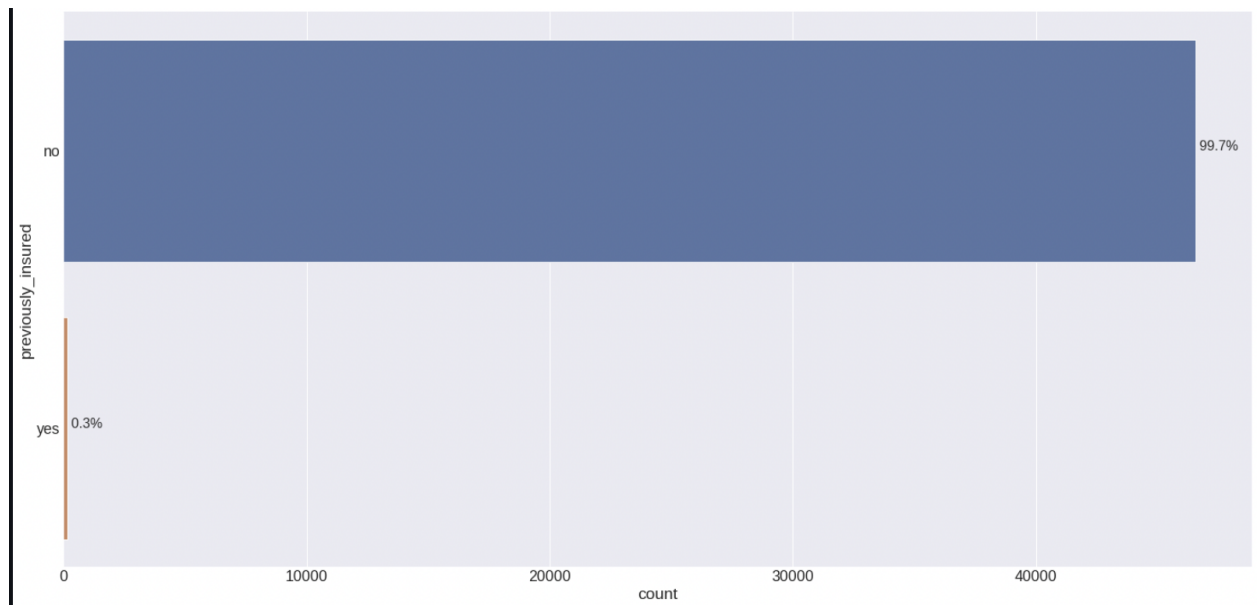
- **A customer's likelihood of wanting health insurance increases with age.**
False: Those who are between the ages of 40 and 50 are more likely to want health insurance.





- **Individuals who own automobile insurance are more likely to purchase health insurance.**

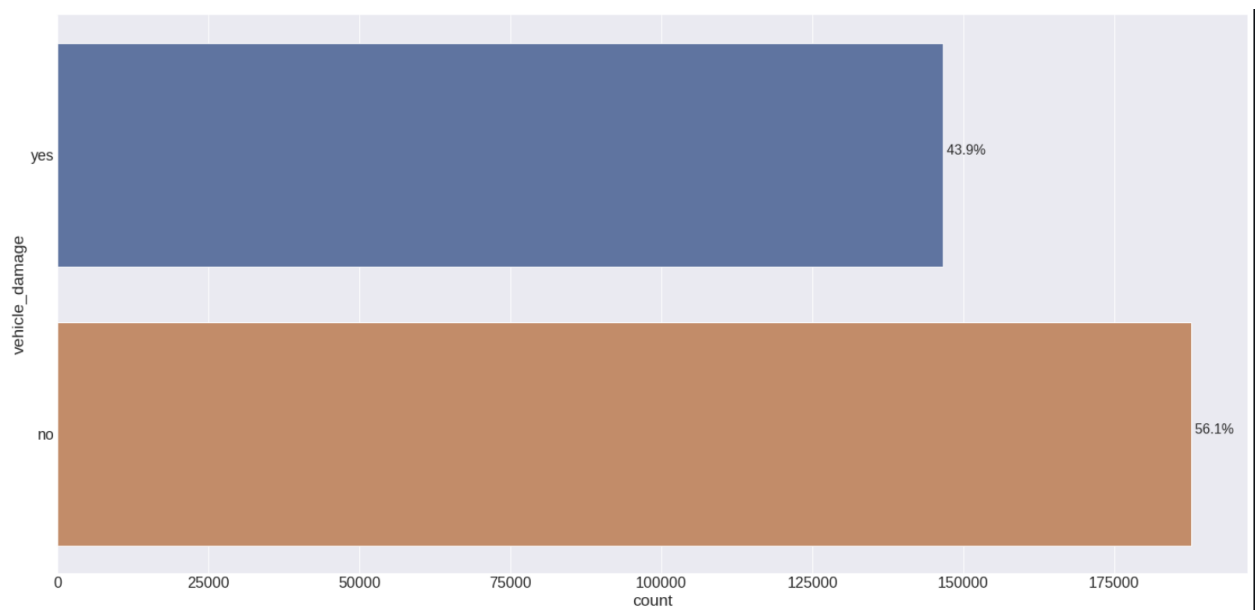
False: People who don't have auto insurance are more likely to have health insurance. Of the customers in this database, 99.7% have had insurance in the past.





- Fewer than 40% of clients who have had their cars damaged do not wish to receive health insurance.

False: About 43.9% of customers are those who don't want health insurance and have suffered damage.





APPLICATION DEPLOYMENT

The demo of the application has been deployed using digital ocean at the url:

<http://67.205.136.29:8000/>

This is a machine learning model that predicts whether a customer interested in Health insurance or Not

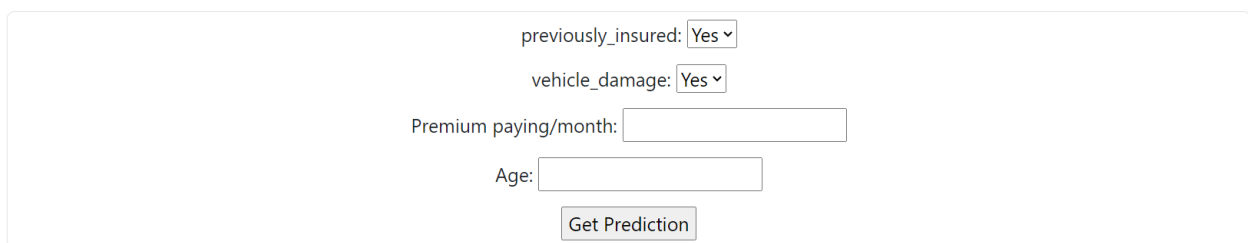
A screenshot of a web application interface for predicting health insurance interest. The form is contained within a light gray rounded rectangle. It features four input fields: two dropdown menus for 'previously_insured' and 'vehicle_damage', both currently set to 'Yes'; a text input field for 'Premium paying/month'; and another text input field for 'Age'. Below these fields is a 'Get Prediction' button with a light gray background and a thin border.

Fig. Application Demo

The code is hosted on GitHub:

<https://github.com/Intro-to-data-science-project/Health-insurance-upsell-with-machine-learning.git>



CONCLUSION

Merely 12.31% +/- 0.15% of the potential customers were accurately classified by the random model. With the ability to distinguish between the classes, the final model was able to classify 24.1% (+/- 0.04%) of cases correctly. Additionally, the lift curve demonstrates that the model achieves a gain three times larger than the customers chosen at random.

Nearly all interested customers (98.31% +/- 0.16%) remain on up to 50% of the sorted list thanks to the model's organization. This allows you to recoup half of the cost of phone calls. Therefore, there would be a cost of R\$ 300,000.00 if each call costs R\$ 15.00 out of 20,000.00. It is feasible to spend just R\$ 150,00.00 by using the model.

Knowledge Acquired

- Accurately interpreting the results and their potential to address business issues is crucial.
- Even a low model score can have an impact on the business issue facing the company.
- I must become knowledgeable about top K scores.



REFERENCES

- [1] <https://vitalflux.com/svm-classifier-scikit-learn-code-examples/>
- [2] <https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction>
- [3] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
- [4] https://www.researchgate.net/publication/370066456_A_Comparative_Analysis_of_Machine_Learning_Models_for_the_Prediction_of_Insurance_Uptake_in_Kenya