Data Analysis Project:
Identifying Future Travel Insurance Customers

Business Case

Business problem

An Indian tour and travels company is offering a Travel Insurance Package to their customers. The new insurance package also includes Covid Cover.

The Insurance was offered to customers in 2019. Data were collected from almost 2000 of their customers and extracted from the performance and sales of of the package during that period. The dataset is available on Kaggle.

The project aims to identify future customers who are interested in buying an insurance package based on historical customer data of the travel agency.

Objectives

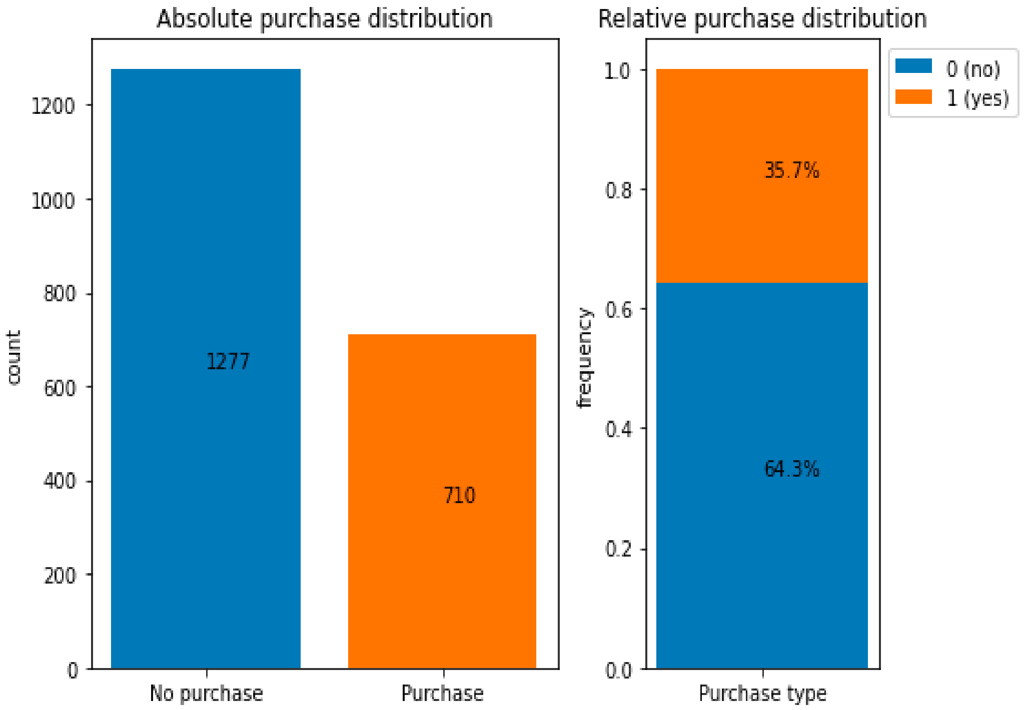
1. To build classification models using different methods.
2. To evaluate the models and choose the most suitable one.
3. To provide the agency useful and practical insights to suggest how to effectively target the right customers.

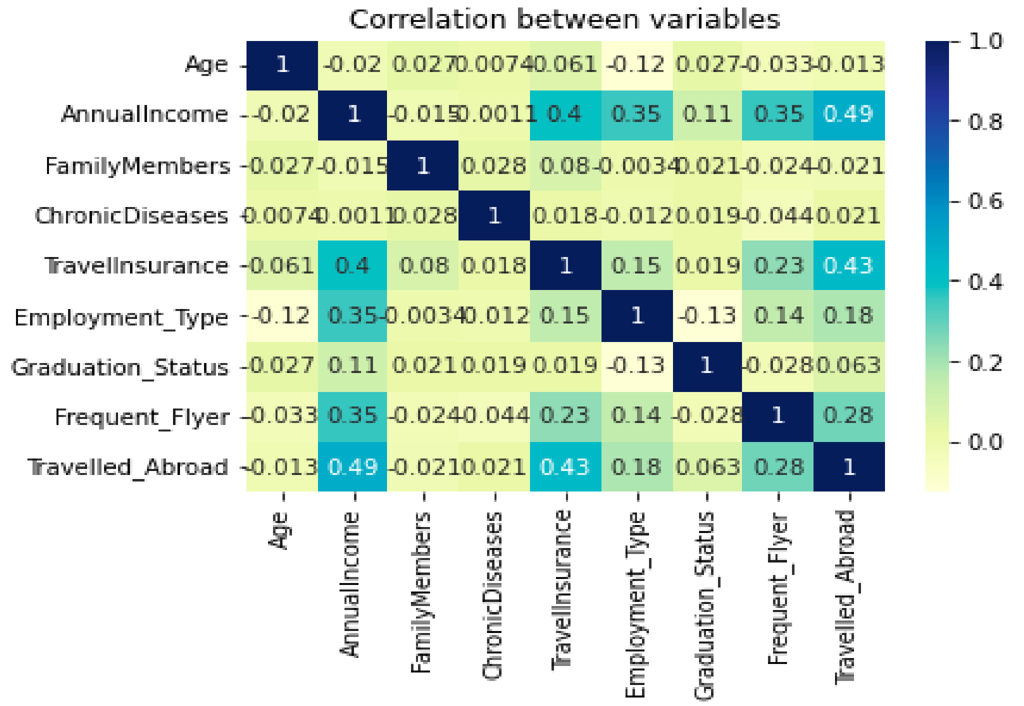
Data Set

- 1987 customers
- 9 variables:
  - 3 non-categorical: Age, AnnualIncome (₹ 10K), FamilyMembers
  - 6 categorical: Employment_Type, Graduation_Status, Frequent_Flyer, Travelled_Abroad, ChronicDiseases, TravelInsurance

	Age	AnnualIncome	FamilyMembers	ChronicDiseases	TravelInsurance	Employment_Type	Graduation_Status	Frequent_Flyer	Travelled_Abroad
0	31	40.0	6	1	0	0	1	0	0
1	31	125.0	7	0	0	1	1	0	0
2	34	50.0	4	1	1	1	1	0	0
3	28	70.0	3	1	0	1	1	0	0
4	28	70.0	8	1	0	1	1	1	0
5	25	115.0	4	0	0	1	0	0	0
6	31	130.0	4	0	0	0	1	0	0
7	31	135.0	3	0	1	1	1	1	1
8	28	145.0	6	1	1	1	1	1	1
9	33	80.0	3	0	0	0	1	1	0

Descriptive Analysis





	count	mean	std	min	25%	50%	75%	max
Age	1987.0	29.650226	2.913308	25.0	28.0	29.0	32.0	35.0
AnnualIncome	1987.0	93.276296	37.685568	30.0	60.0	90.0	125.0	180.0
FamilyMembers	1987.0	4.752894	1.609650	2.0	4.0	5.0	6.0	9.0
ChronicDiseases	1987.0	0.277806	0.448030	0.0	0.0	0.0	1.0	1.0
TravelInsurance	1987.0	0.357323	0.479332	0.0	0.0	0.0	1.0	1.0
Employment_Type	1987.0	0.713135	0.452412	0.0	0.0	1.0	1.0	1.0
Graduation_Status	1987.0	0.851535	0.355650	0.0	1.0	1.0	1.0	1.0
Frequent_Flyer	1987.0	0.209864	0.407314	0.0	0.0	0.0	0.0	1.0
Travelled_Abroad	1987.0	0.191243	0.393379	0.0	0.0	0.0	0.0	1.0

Data Preparation

1. Drop unnecessary column(s)
2. Check any missing values
3. Encode categorical variables
4. Split train to test data 70:30

Model 1 - Logistic Regression

```
                           Logit Regression Results
==============================================================================
Dep. Variable:        TravelInsurance   No. Observations:                 1987
Model:                          Logit   Df Residuals:                     1978
Method:                           MLE   Df Model:                            8
Date:                Mon, 17 Oct 2022   Pseudo R-squ.:                  0.2016
Time:                        15:36:26   Log-Likelihood:                -1034.2
converged:                       True   LL-Null:                       -1295.3
Covariance Type:            nonrobust   LLR p-value:                 1.219e-107
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const               -5.4047      0.634     -8.525      0.000      -6.647      -4.162
Age                  0.0733      0.019      3.958      0.000       0.037       0.110
Employment_Type      0.0986      0.133      0.743      0.457      -0.161       0.358
Graduation_Status   -0.1813      0.156     -1.160      0.246      -0.488       0.125
AnnualIncome         0.0156      0.002      8.844      0.000       0.012       0.019
FamilyMembers        0.1529      0.034      4.551      0.000       0.087       0.219
ChronicDiseases      0.0900      0.121      0.743      0.457      -0.147       0.327
Frequent_Flyer       0.4595      0.137      3.366      0.001       0.192       0.727
Travelled_Abroad     1.7176      0.153     11.211      0.000       1.417       2.018
====================================================================================
```

Statistically significant variables:

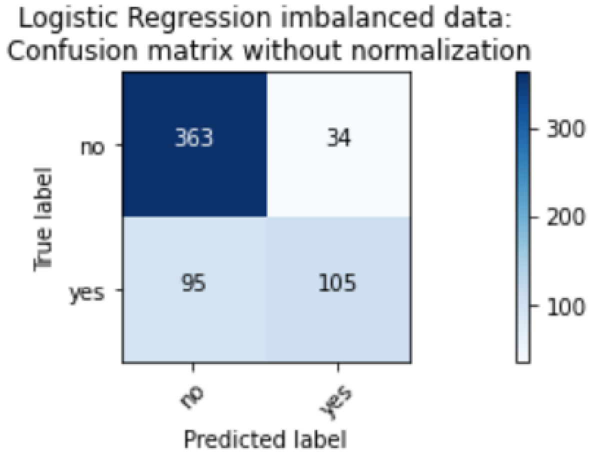
- Travelled_Abroad (1.72)
- Frequent_Flyer (0.46)
- FamilyMembers (0.15)
- Age (0.07)
- AnnualIncome (0.016)

Interpreting the coefficients:

- Travelled_Abroad: $e^{(1.72)} \approx 5.58$ → Customer who has travelled abroad before has 5.58 times the odd of having the travel insurance compared to customer who has not travelled abroad.
- Frequent_Flyer (0.46): $e^{(0.46)} \approx 1.58$
- FamilyMembers (0.15): $e^{(0.15)} \approx 1.16$ → Increase of 1 member in the number of family members increases the odds of having the travel insurance by 16%.
- Age (0.07): $e^{(0.07)} \approx 1.07$
- AnnualIncome (0.016): $e^{(0.16)} \approx 1.016$ → Increase of 10 000 rupees in annual income increases the odds of having the travel insurance by 1.6%.

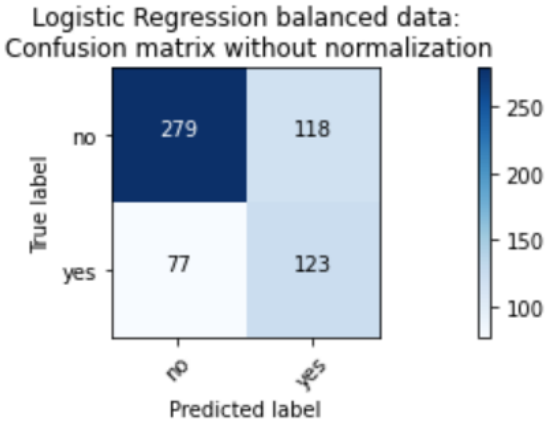
Model 1 - Logistic Regression

Imbalanced Model



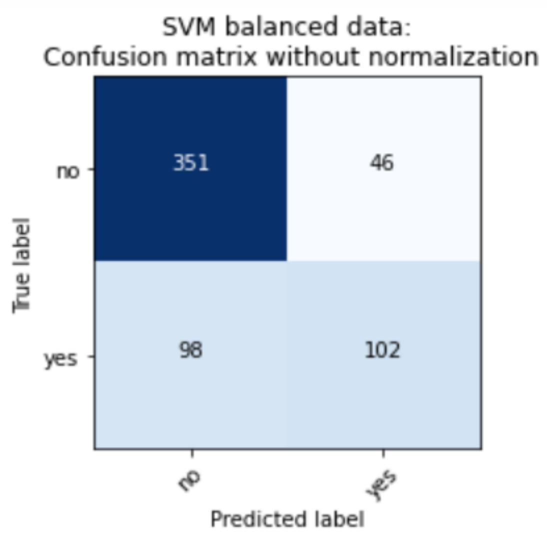
- 2/3 of all customers haven’t bought the insurance, therefore imbalanced model roughly follows this ratio by classifying 3/4 of customers to negative class and 1/4 to positive class.
- Model classifies true negatives accurately.
- Low number of false positives.

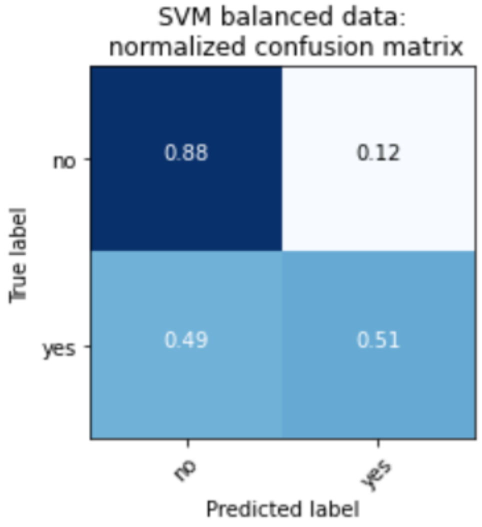
Balanced Model

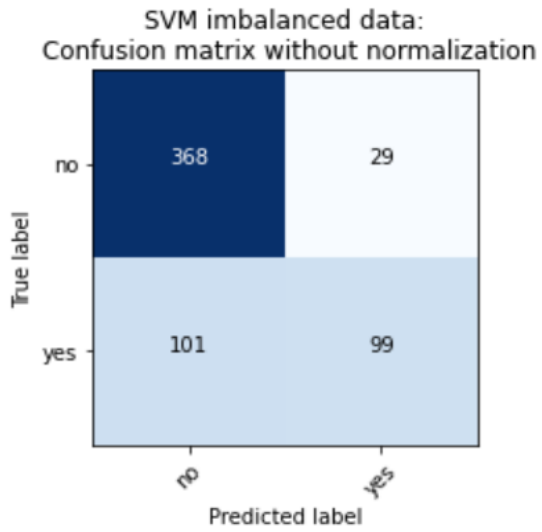


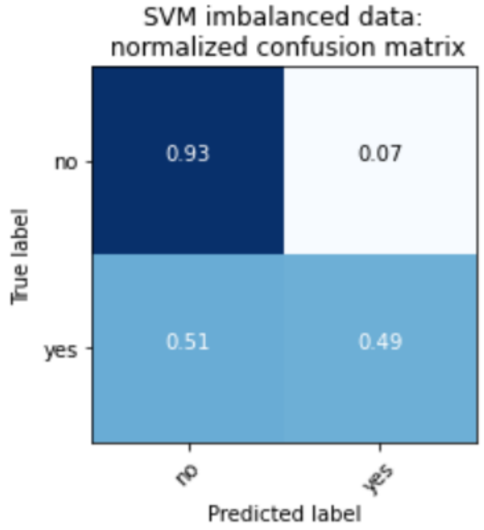
- Balancing the data causes the model to also balance the ratio between negative and positive classifications.
- 60% of predictions were classified as negative and 40% as positive.
- Number of true positives is improved compared to the imbalanced model
- Number of true negatives is much lower and number of false positives is much higher than in the imbalanced model.

Model 2 – Support Vector





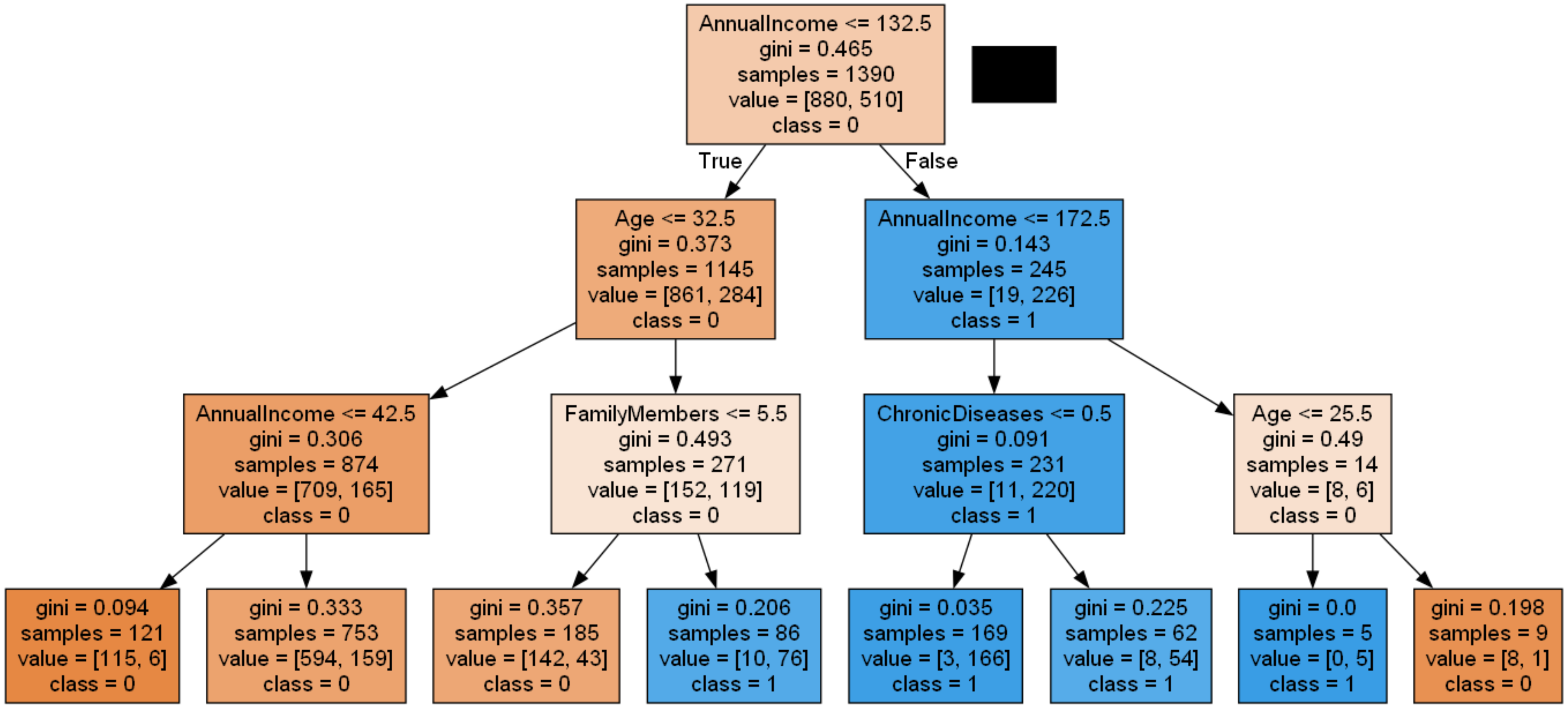


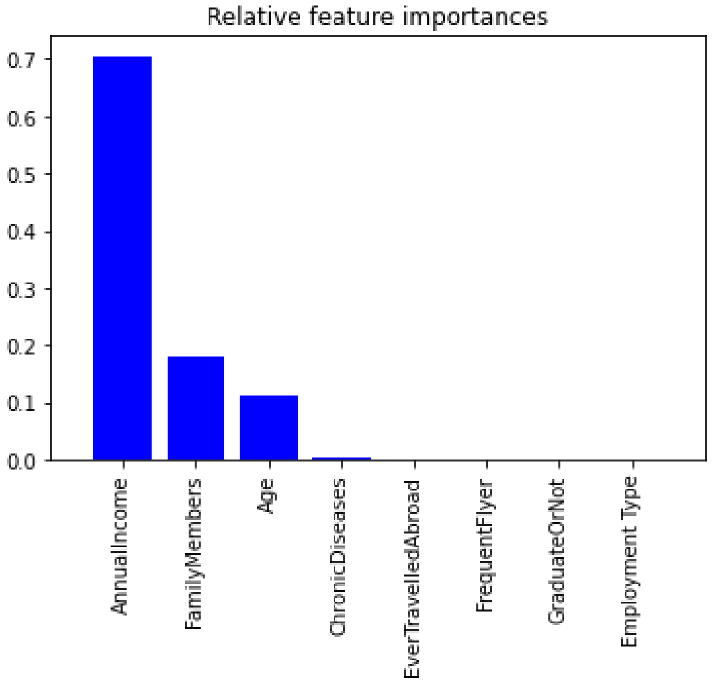


Support Vector Model

- Balanced model has a better True Positive Rate but a worse True Negative Rate than the unbalanced model.
- Unbalanced model is more useful to filter out the people who do not intend to buy the package.
- Balanced model is more accurate in finding the right customers.

Model 3 – Decision Tree

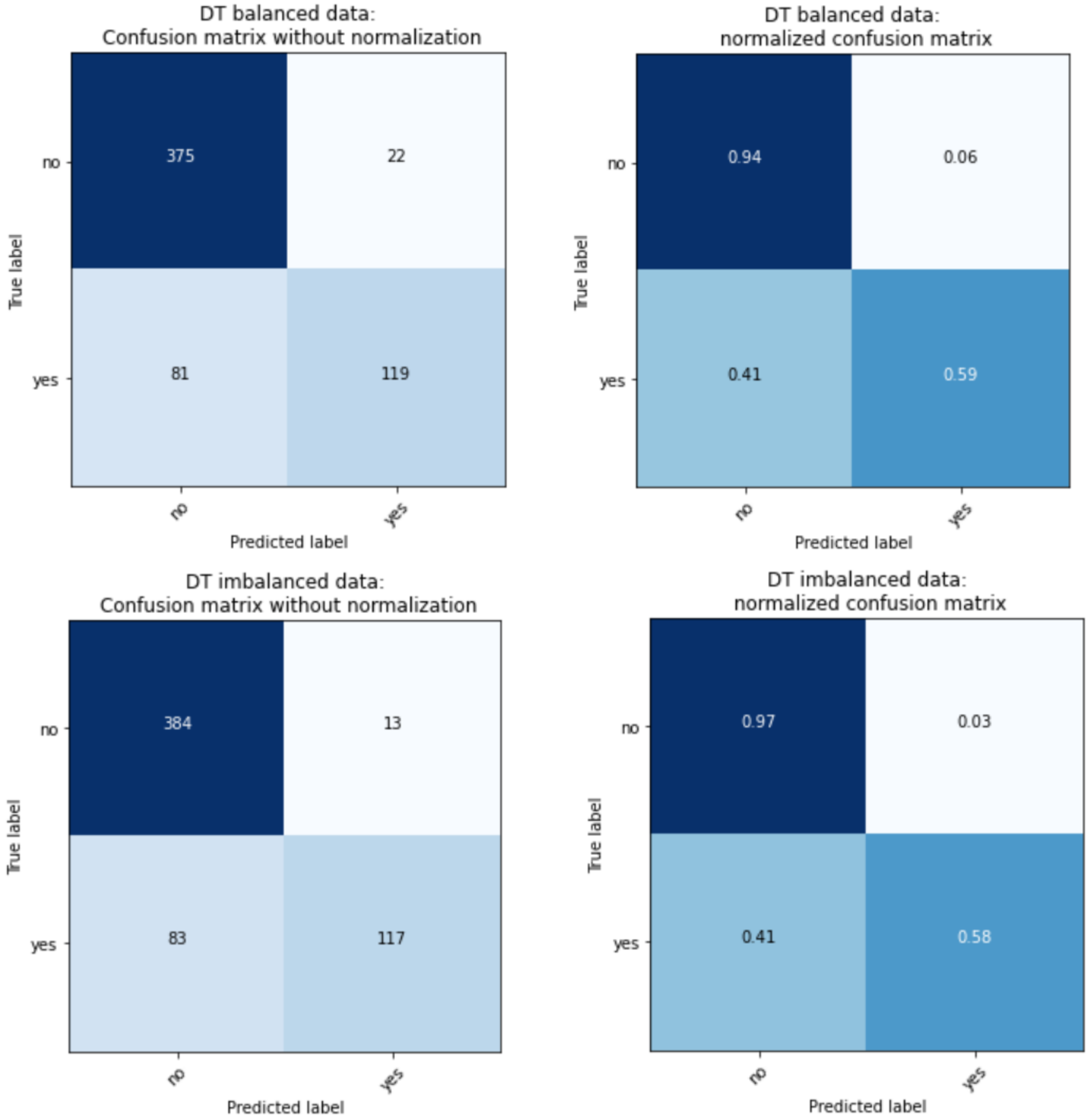




The classifier correctly identifies customers who purchased a travel insurance with 84% of the observations.

According to the unbalanced Decision Tree model, the significant features are annual income, family members and age. Chronic diseases are not seen as a major factor among customers.

Model 3 – Decision Tree



Balanced vs. Unbalanced Data Model

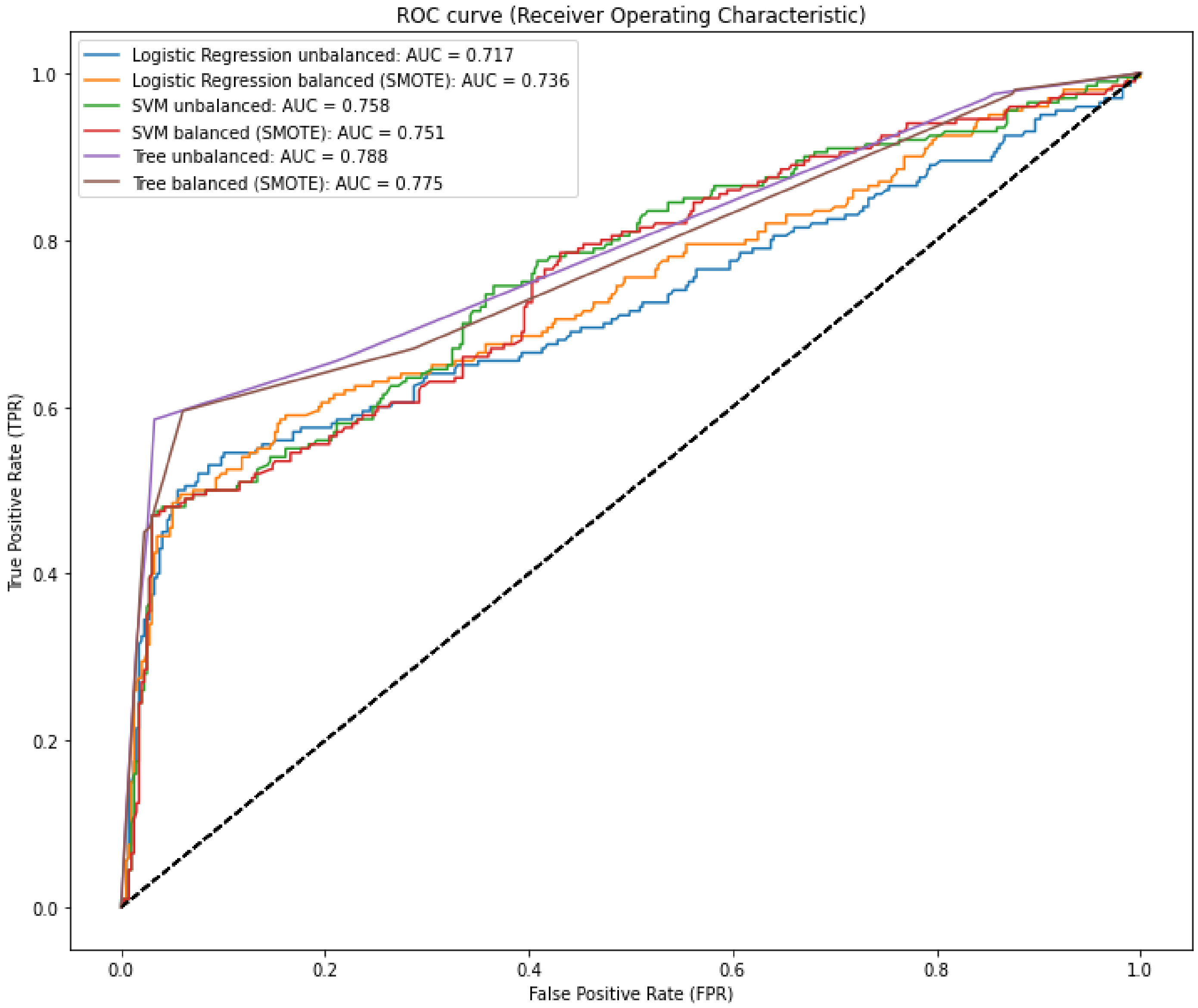
- The balanced data model does not perform far from the unbalanced data model.
- In fact, the unbalanced data model gives a slightly higher score in accuracy, precision, F1 and AUC than the balanced data model.

--> The unbalanced data model shows high reliability.

Model Evaluation 1: Key Statistics

	Logistic Regression		Support-vector machine		Decision Tree	
	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced
True Positive	105	129	99	102	117	119
False Positive	34	114	29	46	13	22
True Negative	363	283	368	351	384	375
False Negative	95	71	101	98	83	81
True Positive Rate	0.53	0.65	0.49	0.51	0.59	0.60
False Positive Rate	0.09	0.29	0.07	0.12	0.03	0.06
Accuracy	0.78	0.69	0.78	0.76	0.84	0.82
Precision	0.76	0.53	0.77	0.69	0.90	0.84
F1 Score	0.62	0.58	0.60	0.59	0.71	0.70
AUC	0.72	0.74	0.76	0.75	0.79	0.78

Model Evaluation 2: ROC curve



The unbalanced Decision Tree appears to be the best performer

Model Evaluation 3: Choose the best

According to key statistics the best model for predicting willingness to buy insurance was **unbalanced decision tree model**.

Strengths vs. other models

- Accuracy 0.84 means that out of all class predictions 84% were correct.
- Precision 0.90 ensures that the model classifies 9/10 positively classified customers correctly into true positives.
- F1 Score 0.71 shows that the model had the highest harmonized mean of precision and true positive rate (recall).
- AUC 0.79 shows that almost 8/10 of class prediction are correct.
- False positive rate 0.03 ensures that only 3% of positive predictions were false

Weaknesses vs. other models

- True positive rate 0.59 was only the third best from all the models. Balanced logistic regression model had the best performance in this metric, but all the other metrics were considerably better in the unbalanced decision tree model.

Recommendations

1. Who are potential customers?

- Individuals with an annual income of between 1,325,000 (₹) and 1,725,000 (₹)
- Couples and families in their 30s who have more than 6 members with an annual income of under 1,325,000 (₹)

2. What should the company do next?

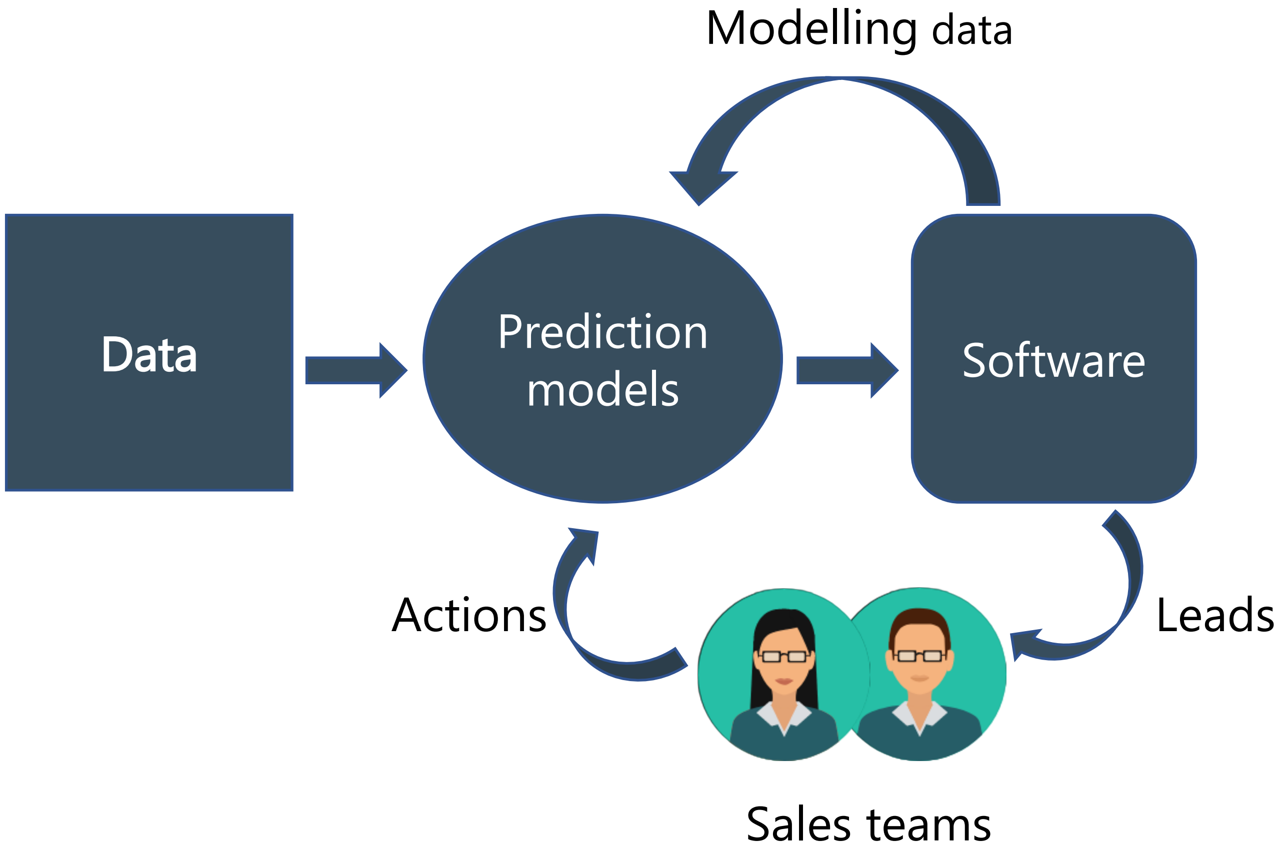
- Better target the right customers with AI tools and ML algorithms
- Use an omnichannel approach to interact with customers
- Cooperate with airline and tourism companies to offer the Travel Insurance with their services and products
- Further analyze why most customers did not buy the package
  -> For example, carry out surveys and interviews
  -> Potentially develope a package that is more catered to different customer groups

"Half the money I spend on advertising is wasted; the trouble is I don't know which half."

- John Wanamaker,

father of modern advertising and a "pioneer in marketing."

Business application: Sales software



1. The application retrieves customer data from the CRM system.
2. The application creates predictions based on customer data with predictive models and ranks the results according to probability.
3. The application gives leads to our sales teams, who record the actions in the application. In this way, it is easy to measure the effectiveness of sales efforts.
4. Actual data about a successful or unsuccessful sale will be returned to the prediction models, which helps the models improve their predictions.