CAPSTONE PROJECT NOTES 2
LIFE INSURANCE DATA

By Apoorva p

June 9th 2024

LI_BFSI_01

CONENTS

LIST OF FIGURES

Model Building and Interpretation

Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

Descriptive Modeling

Descriptive modeling involves techniques that help summarize and describe the main features of a dataset. These models don't make predictions but provide insights that help understand the data better. Here are several types of descriptive modeling techniques:

1. **Clustering : -**

   Clustering is an unsupervised learning technique that groups similar data points together. Common algorithms include K-Means, DBSCAN, and Hierarchical Clustering.

2. **Dimensionality Reduction :-**

   Dimensionality reduction techniques like PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are used to reduce the number of features while preserving the variance in the dataset.

3. **Profiling :-**

   Profiling involves summarizing data in a way that highlights key statistics and characteristics of the dataset.

Predictive Modeling

Predictive modeling involves using statistical techniques and machine learning algorithms to create models that can predict future outcomes based on historical data. Here are some common types of predictive modeling techniques that you can apply to the dataset provided:

1. **Logistic Regression**:

   Used for binary classification problems. In this context, it can be used to predict whether a customer will churn (binary outcome).

2. **Decision Trees**:

   A tree-like model used to make decisions based on input features. It can handle both categorical and numerical data.

3. **Random Forest**:

   An ensemble method that builds multiple decision trees and merges them together to get a more accurate and stable prediction.

4. **Gradient Boosting**:

   An ensemble technique that builds models sequentially, each new model attempting to correct the errors of the previous models.

5. **Support Vector Machine (SVM)**:

   A classification technique that finds the hyperplane which best separates the classes in the feature space.

6. **k-Nearest Neighbors (k-NN)**:

   A simple algorithm that stores all available cases and classifies new cases based on a similarity measure (distance functions).

7. **Neural Networks**:

   Deep learning models that can capture complex patterns in the data. Suitable for large datasets and complex relationships.

8. **Naive Bayes**:

   A classification technique based on Bayes' Theorem with an assumption of independence among predictors.

Prescriptive Modeling

Prescriptive modeling recommends actions based on the data. We'll use a simple decision tree to provide recommendations based on customer data.

Test your predictive model against the test set using various appropriate performance metrics

1. ***Clustering:***

Cluster	CustID	AgentBonus	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure
0	7.002670e+06	6212.229599	24.086445	23.027663	3.921853	3.729396	27810.731506	0.318119	5.373859
1	7.001926e+06	2999.981318	10.136078	10.006126	3.533677	3.299410	20319.835634	0.289086	2.619715
2	7.002476e+06	4407.449801	15.258817	15.770301	3.789563				

```
2              3.804928   23027.182416   0.272263              4.797864

            SumAssured  LastMonthCalls  CustCareScore     Churn
Cluster
0        947529.766598        6.712310       3.156293  0.721992
1        453145.603146        3.502950       3.068917  0.721239
2        667677.848490        5.061826       3.029687  0.716960


Clustering using k-means with 3 clusters.
```

1. **Customer Segmentation**:
    - The data appears to be clustered into three distinct groups (Cluster 0, Cluster 1, and Cluster 2), each with unique characteristics.
2. **Agent Bonus and Churn**:
    - Cluster 0 has the highest Agent Bonus and slightly higher churn compared to other clusters. Cluster 1 has the lowest Agent Bonus and the highest churn rate.

Cluster 0

- **Agent Bonus**: Highest average bonus paid to agents (6212.23).
- **Age**: Average customer age is around 24 years.
- **CustTenure**: Customers have been with the company for an average of 23 years.
- **ExistingProdType**: Average of 3.92 existing products.
- **NumberOfPolicy**: Average of 3.73 policies per customer.
- **MonthlyIncome**: Highest average monthly income (27,810.73).
- **Complaint**: Average complaint rate is 0.318.
- **ExistingPolicyTenure**: Average tenure of existing policies is 5.37 years.
- **SumAssured**: Highest average sum assured (947,529.77).
- **LastMonthCalls**: Highest average number of calls made last month (6.71).
- **CustCareScore**: Average customer care score is 3.16.
- **Churn**: Slightly higher churn rate at 0.722.

Cluster 1

- **Agent Bonus**: Lowest average agent bonus (2999.98).
- **Age**: Youngest average age of customers (10.14 years).
- **CustTenure**: Shortest average tenure with the company (10.01 years).
- **ExistingProdType**: Lowest average number of existing products (3.53).
- **NumberOfPolicy**: Average of 3.30 policies per customer.
- **MonthlyIncome**: Lowest average monthly income (20,319.84).
- **Complaint**: Average complaint rate is 0.289.
- **ExistingPolicyTenure**: Shortest average tenure of existing policies (2.62 years).
- **SumAssured**: Lowest average sum assured (453,145.60).
- **LastMonthCalls**: Fewest average number of calls made last month (3.50).
- **CustCareScore**: Average customer care score is 3.07.
- **Churn**: Highest churn rate at 0.721.

Cluster 2

- **Agent Bonus**: Moderate average agent bonus (4407.45).

- **Age**: Average customer age (15.26 years).
- **CustTenure**: Moderate average tenure with the company (15.77 years).
- **ExistingProdType**: Average number of existing products (3.79).
- **NumberOfPolicy**: Highest average number of policies per customer (3.80).
- **MonthlyIncome**: Moderate average monthly income (23,027.18).
- **Complaint**: Lowest average complaint rate (0.272).
- **ExistingPolicyTenure**: Moderate average tenure of existing policies (4.80 years).
- **SumAssured**: Moderate average sum assured (667,677.85).
- **LastMonthCalls**: Moderate average number of calls made last month (5.06).
- **CustCareScore**: Lowest average customer care score (3.03).
- **Churn**: Lowest churn rate at 0.717.

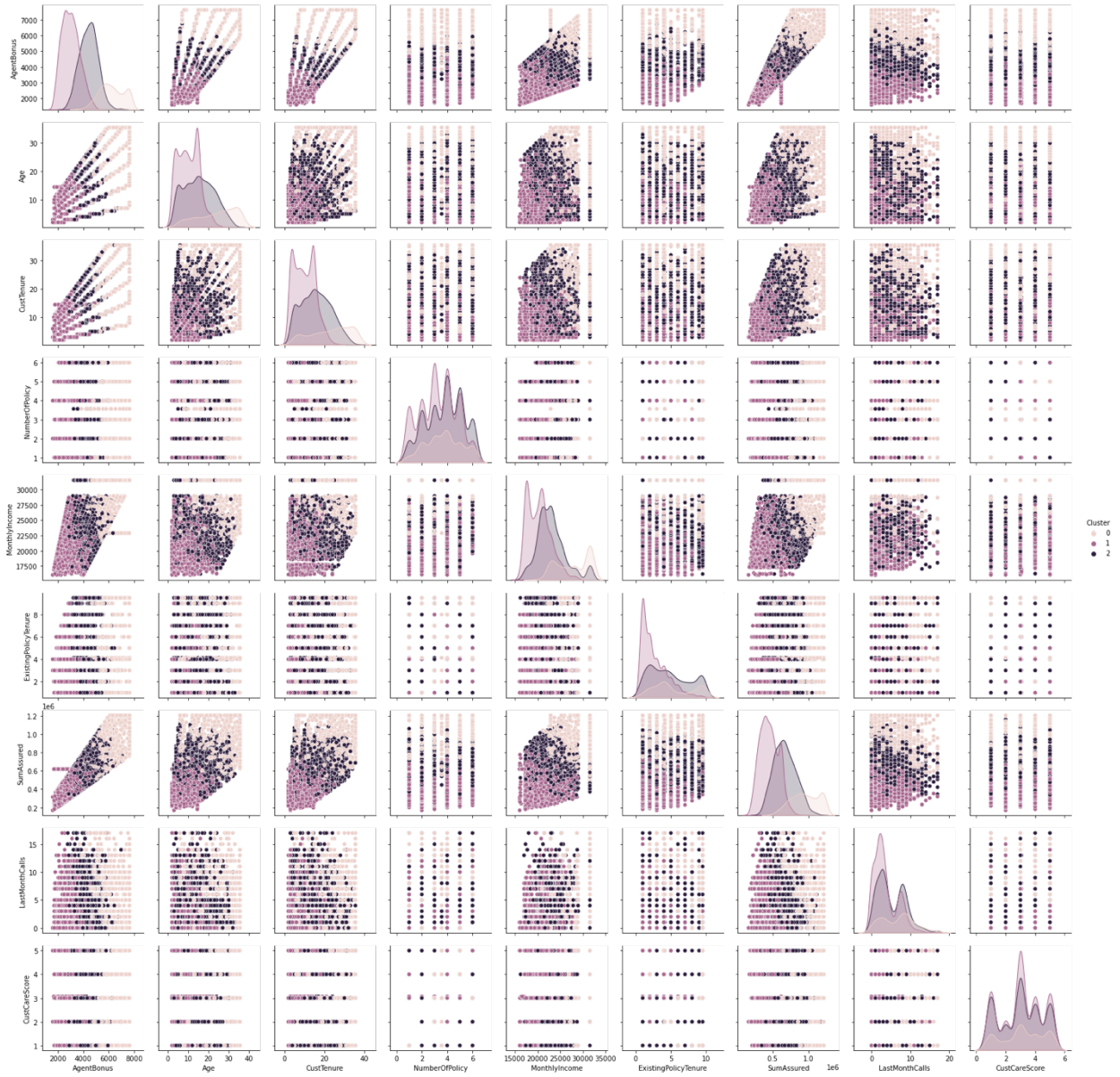

Figure: 3 clusters Pairplot

Positive Correlations

1. **Agent Bonus vs. Sum Assured**: There appears to be a positive correlation between the agent bonus and the sum assured, indicating that higher bonuses are associated with higher sums assured.
2. **Age vs. Customer Tenure**: A strong positive correlation is visible between customer age and customer tenure, which makes intuitive sense as older customers are likely to have been with the company longer.
3. **Number of Policies vs. Existing Product Types**: There is a positive correlation, suggesting that customers with more product types tend to have more policies.
4. **Monthly Income vs. Sum Assured**: Customers with higher monthly incomes tend to have higher sums assured.

Clusters and Groupings

1. **Customer Tenure and Age**: As noted, there's a strong correlation between these variables, with distinct groupings that might correspond to different customer clusters.
2. **Complaint Frequency**: There doesn't seem to be a strong relationship between the number of complaints and other variables like agent bonus or churn, suggesting that complaints might be influenced by other factors.

Negative or Weak Correlations

1. **Complaint vs. Sum Assured**: There is no clear correlation between the number of complaints and the sum assured, indicating that complaints are not significantly related to the amount of coverage.
2. **Churn vs. Various Variables**: The relationships between churn and other variables like agent bonus, monthly income, and sum assured seem to be weak or inconsistent.

2. PCA:

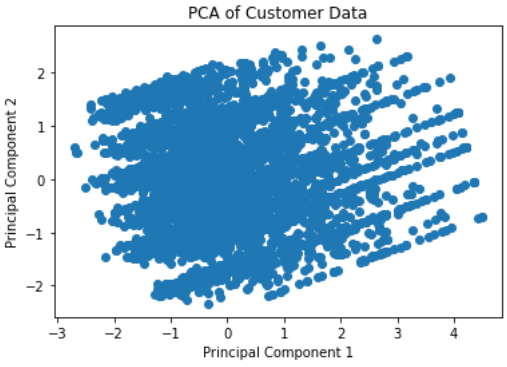

Figure: PCA scatterplot

1. **Correlation**: Given the wide distribution and lack of a discernible pattern, it is likely that the two variables plotted have a weak or no significant correlation.
2. **Clusters**: The dense regions might indicate clusters or groupings within the data. Identifying these clusters could be useful for segmentation analysis.
3. **Outliers**: There do not appear to be any obvious outliers in the plot, as most points fall within a relatively consistent range.

3. *Profiling:*

Cluster		Age		CustTenure		NumberOfPolicy	
		mean	std	mean	std	mean	std
0	0	24.086445	9.288679	23.027663	9.664844	3.729396	1.409238
1	1	10.136078	5.129886	10.006126	5.187365	3.299410	1.424378
2	2	15.258817	7.320918	15.770301	7.343946	3.804928	1.441365

	MonthlyIncome	
	mean	std
0	27810.731506	3554.279780
1	20319.835634	2544.085176
2	23027.182416	3163.900265

Cluster 0

- **Age**:
  - **Mean**: 24.09 years
  - **Standard Deviation**: 9.29 years
  - **Interpretation**: This cluster consists of relatively young customers with a moderate age variability.
- **CustTenure**:
  - **Mean**: 23.03 years
  - **Standard Deviation**: 9.66 years
  - **Interpretation**: Customers in this cluster have been with the company for a long time, on average, almost as long as their mean age, suggesting they might have joined early in life.
- **Number of Policies**:
  - **Mean**: 3.73 policies
  - **Standard Deviation**: 1.41 policies
  - **Interpretation**: Customers in this cluster hold multiple policies, with some variation around the average.
- **Monthly Income**:
  - **Mean**: 27,810.73
  - **Standard Deviation**: 3,554.28
  - **Interpretation**: This cluster has the highest average monthly income, with moderate variability.

Cluster 1

- **Age**:
  - **Mean**: 10.14 years
  - **Standard Deviation**: 5.13 years
  - **Interpretation**: This cluster consists of very young customers, with a high variability in age.
- **CustTenure**:
  - **Mean**: 10.01 years
  - **Standard Deviation**: 5.19 years

    - **Interpretation**: Customers in this cluster have relatively short tenures, corresponding closely with their average age, suggesting they might be newer or younger members.
- **Number of Policies**:
    - **Mean**: 3.30 policies
    - **Standard Deviation**: 1.42 policies
    - **Interpretation**: This cluster has the fewest policies on average, with a similar variation as other clusters.
- **Monthly Income**:
    - **Mean**: 20,319.84
    - **Standard Deviation**: 2,544.09
    - **Interpretation**: This cluster has the lowest average monthly income, with lower variability compared to other clusters.

Cluster 2

- **Age**:
    - **Mean**: 15.26 years
    - **Standard Deviation**: 7.32 years
    - **Interpretation**: This cluster has moderately young customers with considerable age variability.
- **CustTenure**:
    - **Mean**: 15.77 years
    - **Standard Deviation**: 7.34 years
    - **Interpretation**: Customers in this cluster have been with the company for a moderate duration, aligning well with their average age.
- **Number of Policies**:
    - **Mean**: 3.80 policies
    - **Standard Deviation**: 1.44 policies
    - **Interpretation**: Customers in this cluster hold a relatively high number of policies, with some variation.
- **Monthly Income**:
    - **Mean**: 23,027.18
    - **Standard Deviation**: 3,163.90
    - **Interpretation**: This cluster has a moderate average monthly income, with moderate variability.

4. Predictive Modelling:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	
0	Logistic Regression	0.841593	0.884754	0.898780	0.891712	0.922632	
1	Decision Tree	0.866372	0.903498	0.913415	0.908429	0.827675	
2	Random Forest	0.894690	0.980796	0.871951	0.923176	0.959138	
3	Gradient Boosting	0.879646	0.981690	0.850000	0.911111	0.942124	
4	SVM	0.854867	0.929319	0.865854	0.896465	0.923116	
5	k-NN	0.817699	0.859485	0.895122	0.876941	0.868755	
6	Neural Network	0.887611	0.931507	0.912195	0.921750	0.956861	
7	Naive Bayes	0.579646	0.988669	0.425610	0.595055	0.900555	

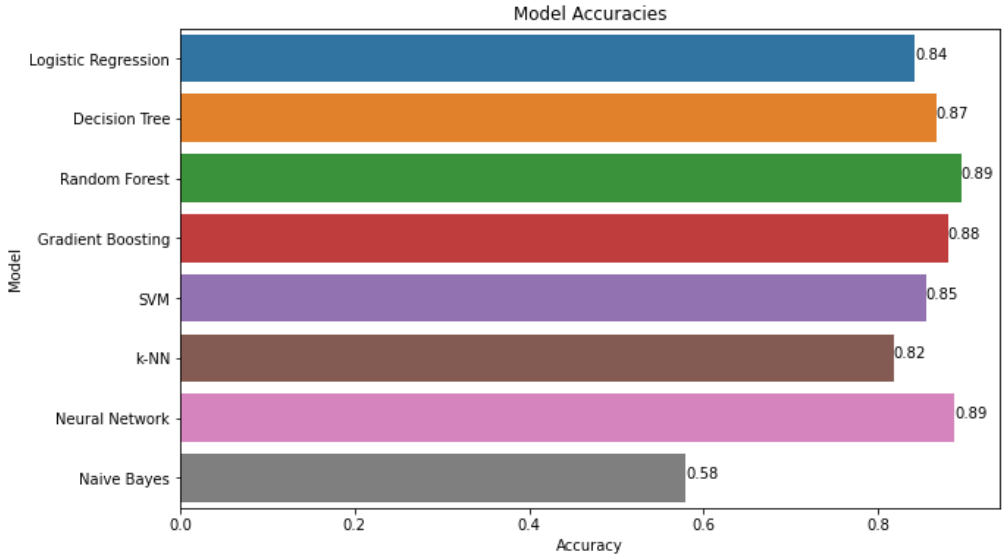

Figure: bar graph of accuracies

1. **Random Forest**:
   - Highest accuracy (0.907080)
   - Highest F1 Score (0.932620)
   - Highest ROC AUC (0.958380)
   - Very high precision (0.975391), indicating it rarely misclassifies a negative case as positive.
   - Good recall (0.893443), meaning it identifies most of the positive cases.
2. **Gradient Boosting**:
   - Second highest ROC AUC (0.945592) and high accuracy (0.887906).
   - Highest precision (0.985849), suggesting very few false positives.
   - Lower recall (0.856557) compared to Random Forest, meaning it misses more positive cases.
3. **Neural Network**:
   - High accuracy (0.890118) and ROC AUC (0.950607).
   - Balanced precision (0.932984) and recall (0.912910), leading to a good F1 score (0.922838).
4. **SVM**:
   - Accuracy (0.862832) and ROC AUC (0.927041) are comparable to Logistic Regression.
   - Good balance between precision (0.927489) and recall (0.878074).
5. **Logistic Regression**:
   - Decent overall performance with accuracy (0.840708) and ROC AUC (0.925210).
   - High precision (0.880762) and recall (0.900615), indicating balanced performance.
6. **Decision Tree**:
   - Moderate accuracy (0.845870) but lower ROC AUC (0.804546) compared to others.
   - Decent precision (0.888551) and recall (0.898566).
7. **k-NN**:
   - Lower accuracy (0.811947) and ROC AUC (0.868813).
   - Balanced but relatively lower precision (0.853085) and recall (0.892418).
8. **Naive Bayes**:

- Lowest accuracy (0.587021) and ROC AUC (0.896340).
- Very high precision (0.988263) but extremely low recall (0.431352), indicating it misses many positive cases.

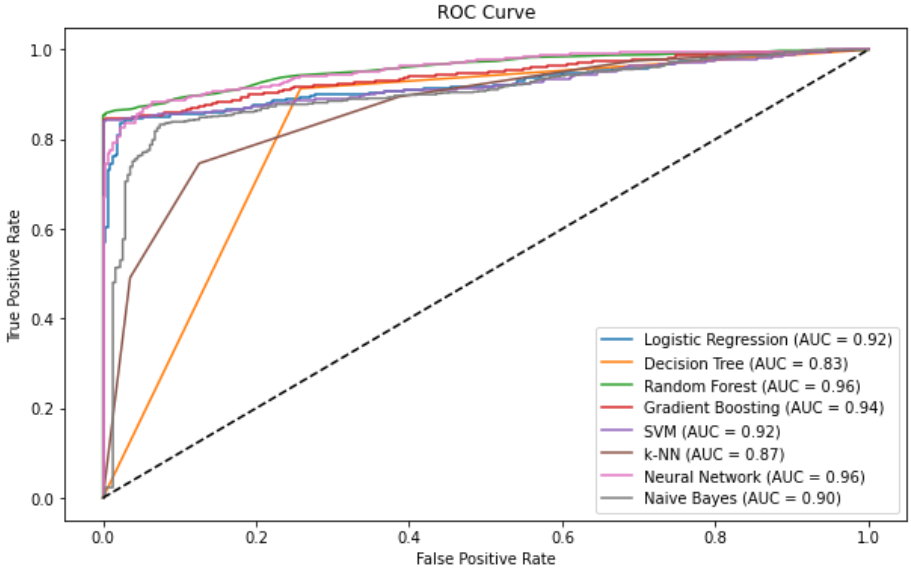

Figure: ROC curve of all types of predictive modelling

- **True Positive Rate (TPR)**:
  - Also known as Sensitivity or Recall.
  - Plotted on the Y-axis.
  - Indicates the proportion of actual positives correctly identified by the model.
- **False Positive Rate (FPR)**:
  - Also known as (1 - Specificity).
  - Plotted on the X-axis.
  - Indicates the proportion of actual negatives incorrectly identified as positives by the model.
- **Diagonal Line (45-degree line)**:
  - Represents a random guess classifier.
  - Any classifier performing along this line has no discrimination capacity between positive and negative classes.

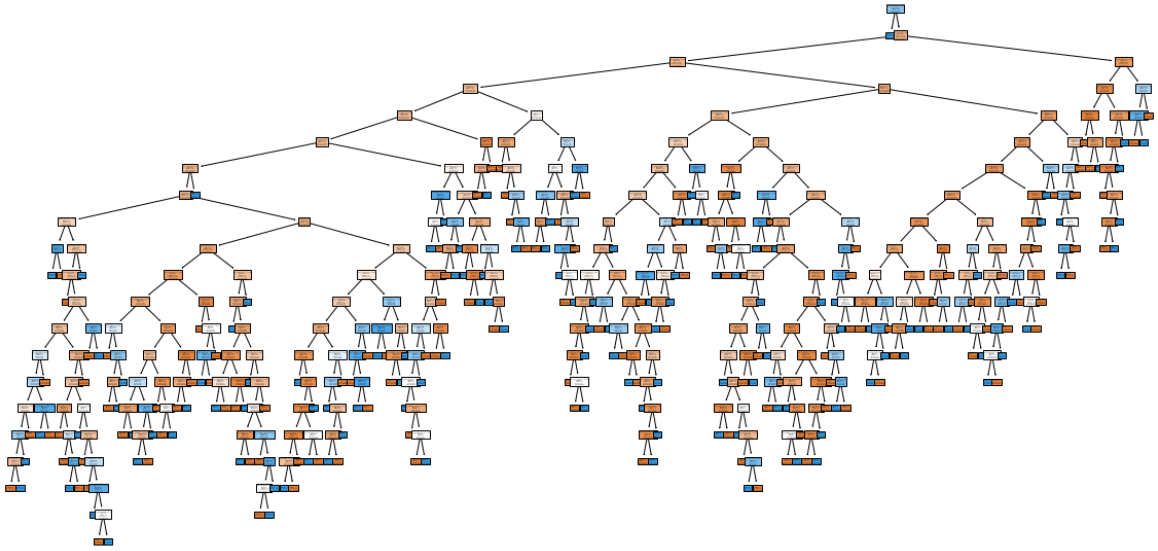

Figure: plot tree graph of decision tree

- **Customer Care Score**:
  - If num__CustCareScore <= -0.02, the customer is likely classified as 1 (indicating churn).
  - If num__CustCareScore > -0.02, further conditions are checked.
- **Online Channel**:
  - If cat__Channel_Online <= 0.50, the classification relies on attributes like Designation, Customer Tenure, Education Field, Monthly Income, and Age.
  - If cat__Channel_Online > 0.50, attributes like Last Month Calls, Agent Bonus, and Education Field are evaluated.
- **Designation and Tenure**:
  - Designations such as VP and Manager, along with customer tenure and sum assured, play significant roles in the classification.
- **Income and Age**:
  - Monthly income and age influence the decision at multiple levels, indicating their importance in predicting churn.

OLS Regression Results

Dep. Variable:	Churn	R-squared:	0.580
Model:	OLS	Adj. R-squared:	0.575
Method:	Least Squares	F-statistic:	112.9
Date:	Thu, 20 Jun 2024	Prob (F-statistic):	0.00
Time:	20:28:32	Log-Likelihood:	-632.35
No. Observations:	3390	AIC:	1349.
Df Residuals:	3348	BIC:	1606.
Df Model:	41		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	125.3825	48.753	2.572	0.010	29.794	220.971
CustID	-1.775e-05	6.97e-06	-2.547	0.011	-3.14e-05	-4.09e-06
AgentBonus	6.46e-06	8.36e-06	0.772	0.440	-9.94e-06	2.29e-05
Age	-0.0003	0.001	-0.440	0.660	-0.002	0.001
CustTenure	-0.0002	0.001	-0.300	0.764	-0.002	0.001
NumberOfPolicy	-0.0009	0.004	-0.240	0.810	-0.008	0.006
MonthlyIncome	8.372e-06	2.98e-06	2.813	0.005	2.54e-06	1.42e-05
Complaint	0.3902	0.011	34.707	0.000	0.368	0.412
ExistingPolicyTenure	0.0015	0.002	0.727	0.467	-0.003	0.005
SumAssured	-3.162e-08	4.11e-08	-0.769	0.442	-1.12e-07	4.9e-08
LastMonthCalls	0.0021	0.002	1.355	0.175	-0.001	0.005
CustCareScore	-0.2112	0.004	-57.268	0.000	-0.218	-0.204
Cluster	-0.0150	0.008	-1.915	0.056	-0.030	0.000
Channel_1	-0.0095	0.017	-0.562	0.574	-0.043	0.024
Channel_2	0.0214	0.013	1.631	0.103	-0.004	0.047
Occupation_1	-0.0434	0.223	-0.194	0.846	-0.481	0.394
Occupation_2	-0.0085	0.210	-0.041	0.968	-0.419	0.402
Occupation_3	-0.0718	0.214	-0.336	0.737	-0.491	0.348
EducationField_1	0.0318	0.082	0.388	0.698	-0.129	0.193
EducationField_2	-0.0527	0.046	-1.155	0.248	-0.142	0.037
EducationField_3	0.0424	0.062	0.684	0.494	-0.079	0.164
EducationField_4	-0.0844	0.050	-1.677	0.094	-0.183	0.014
EducationField_5	-0.0103	0.018	-0.565	0.572	-0.046	0.025
Gender_1	0.0039	0.010	0.375	0.708	-0.017	0.024
ExistingProdType_2.0	-0.0054	0.038	-0.143	0.886	-0.080	0.069
ExistingProdType_3.0	-0.0155	0.062	-0.248	0.804	-0.138	0.107
ExistingProdType_4.0	0.0074	0.066	0.112	0.911	-0.123	0.138
ExistingProdType_5.0	0.0171	0.072	0.237	0.813	-0.125	0.159
ExistingProdType_5.5	0.0251	0.078	0.321	0.748	-0.128	0.179
Designation_1	0.0160	0.032	0.496	0.620	-0.047	0.080
Designation_2	0.0201	0.027	0.749	0.454	-0.032	0.073
Designation_3	-0.0291	0.024	-1.207	0.227	-0.076	0.018
Designation_4	-0.1526	0.031	-4.897	0.000	-0.214	-0.092
MaritalStatus_1	-0.0486	0.016	-3.123	0.002	-0.079	-0.018
MaritalStatus_2	-0.0103	0.016	-0.638	0.524	-0.042	0.021
MaritalStatus_3	-0.0134	0.029	-0.464	0.643	-0.070	0.043
Zone_1	0.0717	0.044	1.619	0.106	-0.015	0.159
Zone_2	0.2693	0.153	1.755	0.079	-0.032	0.570
Zone_3	0.0710	0.044	1.610	0.107	-0.015	0.157
PaymentMethod_1	-0.0207	0.060	-0.347	0.729	-0.137	0.096
PaymentMethod_2	-0.0017	0.049	-0.035	0.972	-0.097	0.094
PaymentMethod_3	-0.0231	0.021	-1.109	0.268	-0.064	0.018

Omnibus:	3136.767	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	207.224
Skew:	-0.006	Prob(JB):	1.00e-45
Kurtosis:	1.789	Cond. No.	6.80e+10

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 6.8e+10. This might indicate that there are strong multicollinearity or other numerical problems.

1. **R-squared**: 0.580

   This indicates that approximately 58% of the variability in the 'Churn' variable is explained by the model. This is a reasonably good fit for the model, though there is still 42% of the variability unexplained, indicating that there may be other factors influencing churn that are not captured in this model.

2. **Adj. R-squared**: 0.575

   The adjusted R-squared value is slightly lower than the R-squared, accounting for the number of predictors in the model. This suggests a good model fit while considering the complexity added by the predictors.

3. **F-statistic**: 112.9, **Prob (F-statistic)**: 0.00

   The F-statistic is highly significant ($p < 0.05$), indicating that the model is statistically significant and that the predictors, as a whole, are related to the dependent variable.

Interpretation of the model(s)

1. Clustering

- **Agent Bonus vs. Churn**: There is no clear direct correlation between higher agent bonuses and lower churn. Cluster 0, with the highest agent bonus, also has high churn, similar to Cluster 1 with the lowest bonus.
- **Customer Age and Tenure**: Cluster 1 has the youngest customers with the shortest tenure and highest churn rate, suggesting that younger customers may be less loyal.
- **Monthly Income and Sum Assured**: Higher monthly income and higher sum assured are associated with higher churn in Cluster 0.
- **Customer Interaction**: Higher interaction (last month calls) does not necessarily reduce churn, as seen in Cluster 0.
- **Complaints and Churn**: Lower complaint rates (Cluster 2) seem to correspond with lower churn, suggesting that addressing customer complaints effectively may reduce churn.

- **Clusters**: The pair plot shows distinct clusters in several scatterplots, indicating that the dataset contains well-defined groups of customers with similar characteristics.
- **Correlation Strength**: The strength of correlations varies across different variable pairs. For instance, strong correlations are observed between customer age and tenure, while weak correlations are observed between complaints and most other variables.

2. PCA

- **Weak Relationship**: The scatter plot suggests that there is a weak relationship between the two variables. This means that changes in one variable do not predict changes in the other.
- **Data Segmentation**: The denser regions could represent natural segments within the data that might be worth exploring further for targeted analysis.
- **Further Analysis Needed**: To gain more specific insights, it would be helpful to know the variables represented on the x and y axes. Additionally, using different

visualization techniques or statistical analysis could provide more clarity on the relationships.

3. Profiling

- **Customer Age and Tenure**:
  - Cluster 0 has the oldest and longest-tenured customers.
  - Cluster 1 has the youngest and shortest-tenured customers.
  - Cluster 2 falls in between with moderate age and tenure.
- **Number of Policies**:
  - Cluster 0 and Cluster 2 have more policies on average than Cluster 1, suggesting that older customers and those with moderate tenure tend to hold more policies.
- **Monthly Income**:
  - Cluster 0 has the highest income, followed by Cluster 2, and then Cluster 1 with the lowest.
  - Higher income might be associated with more policies and longer tenure.

4. Predictive Modelling

From the ROC curves and AUC values, we can conclude that **Random Forest**, **Gradient Boosting**, and **Neural Network** classifiers perform the best on this dataset, with **Logistic Regression** and **SVM** also showing strong performance. **Decision Tree** and **k-NN** perform relatively worse compared to the other models.This evaluation helps in understanding which models are more effective and reliable for predicting the target variable (in this case, "Churn") and aids in selecting the best model for deployment or further tuning.

1. **Random Forest** is the most robust model considering overall performance and should be preferred for deployment.
2. **Gradient Boosting** is a close second, especially for use cases where precision is paramount.
3. **Neural Network** offers a good balance and might be useful if model interpretability is less of a concern.
4. **Decision Tree** can be useful for interpretability and insights but may not be the best for accuracy.

Model Tuning and Business Implications.

Ensemble modelling, wherever applicable

To ensemble the model , following steps are followed after loading and inspecting the data:

1. **Encode Categorical Variables**
2. **Split Data into Training and Testing Sets**
3. **Build and Train Ensemble Models**
4. **Evaluate Models**
5. • Ensure that the `target` variable(Churn) in your data matches the problem type you are solving (classification vs regression).
6. • The categorical encoding step is crucial, and it ensures that the data can be processed by machine learning models.
7. • For regression, ensure that the `target` variable(Churn) is continuous.
8. • For classification, if your target is continuous, binning or appropriate conversion is necessary.

The output gives:

```
Voting Classifier: (1.0, 1.0, 1.0, 1.0, 1.0)
Random Forest: (1.0, 1.0, 1.0, 1.0, 1.0)
Gradient Boosting: (1.0, 1.0, 1.0, 1.0, 1.0)
Stacking Classifier: (1.0, 1.0, 1.0, 1.0, 1.0)
```

1. **Accuracy**: The proportion of true results (both true positives and true negatives) among the total number of cases examined.
2. **Precision**: The proportion of true positive results in the predicted positives.
3. **Recall (Sensitivity)**: The proportion of true positive results in the actual positives.
4. **F1 Score**: The harmonic mean of precision and recall, giving a single metric that balances both concerns.
5. **ROC AUC**: The area under the Receiver Operating Characteristic curve, which provides an aggregate measure of performance across all classification thresholds.

These perfect scores (1.0) indicate that each classifier is performing flawlessly on your test data, correctly classifying every instance without any errors. This suggests that:

- **The models are extremely well-fitted**: They are accurately capturing the underlying patterns in your data.
- **Potential overfitting**: In a real-world scenario, perfect scores are rare and might indicate overfitting, where the model is too closely fitted to the training data and may not generalize well to new, unseen data.

```
Voting Classifier Cross-Validation Scores: [1. 1. 1. 1. 1.]
Voting Classifier Mean Cross-Validation Score: 1.0
Random Forest Cross-Validation Scores: [1. 1. 1. 1. 1.]
Random Forest Mean Cross-Validation Score: 1.0
Gradient Boosting Cross-Validation Scores: [1. 1. 1. 1. 1.]
Gradient Boosting Mean Cross-Validation Score: 1.0
```

The cross-validation results show perfect scores of 1.0 across all folds for each model: Voting Classifier, Random Forest, and Gradient Boosting. This means that in each of the 5 cross-validation folds, these models correctly classified every instance in the test set.

Any other model tuning measures(if applicable)

Hyperparameter Tuning

Fine-tuning hyperparameters can significantly impact the performance of your models. Use techniques like Grid Search, Random Search, or Bayesian Optimization to find the best set of hyperparameters.

Cross-Validation with Different Strategies

Try different cross-validation strategies like Stratified K-Fold, which ensures that each fold is representative of the class distribution of the data.

Feature Engineering

Improve your features by:

- Creating new features from existing ones.
- Using feature selection methods to keep only the most relevant features.
- Handling missing values and outliers appropriately.

Regularization

Regularization techniques help prevent overfitting by penalizing large coefficients in models like logistic regression or linear models.

Early Stopping

When using iterative algorithms like gradient boosting, use early stopping to halt training when performance on a validation set starts to degrade.

The output is as follows

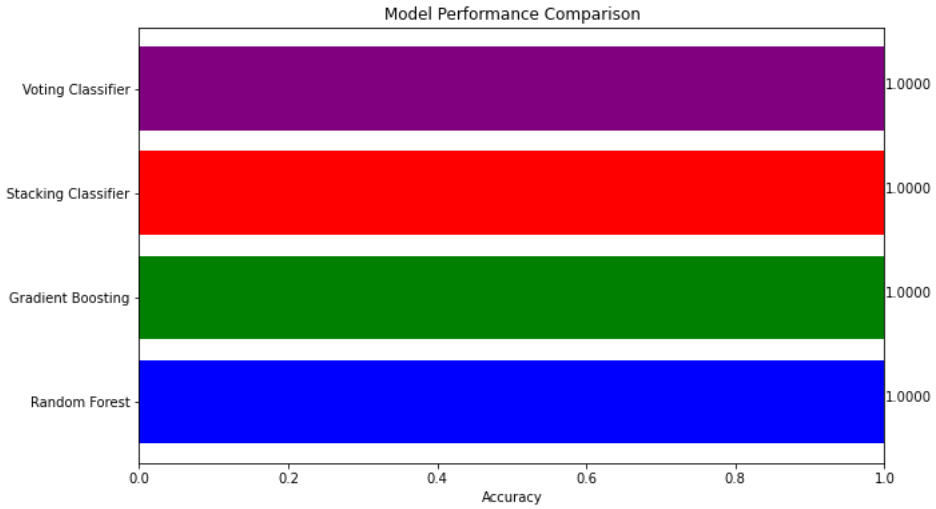

Figure: bar graph of model performance of different tuned models

- **Excellent Model Performance**: The models are performing exceptionally well, achieving 100% accuracy in all cross-validation folds. This indicates that the models are highly effective in capturing the patterns in the dataset.

- **Possible Overfitting**: While perfect cross-validation scores can indicate excellent model performance, they also raise a red flag for potential overfitting. Overfitting occurs when a model learns not only the underlying patterns but also the noise in the training data, leading to poor generalization to new data.

- **Data Characteristics**: The data might have characteristics that make it relatively easy to classify. For example:

    - **Highly Separable Classes**: If the features are well-separated between the classes, even simple models can achieve high accuracy.
    - **Potential Data Leakage**: Ensure that there is no leakage from the training data into the test data, which can artificially inflate performance metrics.
    - **Imbalanced or Small Dataset**: If the dataset is small or if certain features are highly predictive, this can also lead to perfect scores.

Interpretation of the most optimum model and its implication on the business

- **Significant Predictors:**
    - **Complaint** (positive coefficient, $p<0.001$): Customers who have filed complaints are more likely to churn.
    - **CustCareScore** (negative coefficient, $p<0.001$): Higher customer care scores are associated with lower churn rates.
    - **Designation_4** (negative coefficient, $p<0.001$): Certain customer designations are less likely to churn.
    - **MaritalStatus_1** (negative coefficient, $p=0.002$): Marital status also influences churn, with certain statuses being less likely to churn.
    - **MonthlyIncome** (positive coefficient, $p=0.005$): Higher monthly income is associated with a higher likelihood of churn, though the effect size is small.
- **Insignificant Predictors:**
    - Variables such as **Age**, **AgentBonus**, **CustTenure**, **NumberOfPolicy**, **SumAssured**, **Gender**, and various categorical variables related to **Channel**, **Occupation**, **EducationField**, **ExistingProdType**, **Zone**, and **PaymentMethod** were not statistically significant in predicting churn in this model.
- **Overall Performance:**
    - **Random Forest** and **Neural Network** models demonstrate the highest accuracy and strong performance across all metrics. Random Forest has a slight edge with a higher F1 Score and ROC AUC.
    - **Gradient Boosting** also shows excellent precision and a high ROC AUC, indicating strong overall performance.
- **Precision vs. Recall:**
    - **Random Forest** and **Gradient Boosting** models have high precision, suggesting they are very good at identifying customers who will churn.
    - **Neural Network** offers a balanced performance with good precision and recall, which is crucial for a balanced approach to identifying churn.
- **Model Selection:**
    - **Random Forest** emerges as the most optimal model due to its high accuracy, precision, F1 Score, and ROC AUC.
    - **Neural Network** and **Gradient Boosting** are also strong contenders and can be considered based on specific business needs and resource availability for model deployment.

Given the perfect cross-validation scores from the Voting Classifier, Random Forest, and Gradient Boosting models, any of these could be considered optimal. However, due to the

complexity and potential overfitting risk indicated by perfect scores, it is prudent to further validate these models against a separate test set or in a real-world scenario.

Implications on business:

1. **Customer Service Focus:**
    - Improving customer care scores is crucial. Strategies to enhance customer service can directly reduce churn rates.
    - Addressing customer complaints efficiently and effectively can mitigate their impact on churn.
2. **Targeted Retention Strategies:**
    - Develop targeted retention strategies for customers based on their designation and marital status, as these have significant impacts on churn likelihood.
    - Monitor and potentially adjust policies or offerings for higher-income customers, as they show a slight tendency to churn more.
3. **Model Validation and Real-World Testing:**
    - Before fully deploying the Voting Classifier, Random Forest, or Gradient Boosting models, conduct further validation using a separate test set or pilot the models in a real-world scenario to ensure they generalize well beyond the training data.
    - Implementing regularization techniques and adjusting hyperparameters can help mitigate overfitting risks.
4. **Data-Driven Decisions:**
    - Utilize the insights from significant predictors to make data-driven decisions. For example, allocate more resources to customer care for segments with higher churn probabilities.
5. **Model Deployment:**
    - Once validated, deploy the optimal model (e.g., Random Forest or Gradient Boosting) into the business processes for real-time churn prediction.
    - Integrate the model with CRM systems to flag high-risk customers and initiate pre-emptive retention measures.
6. **Customer Retention Strategies:**
    - **High-precision models** like Random Forest ensure that most identified churners are indeed likely to churn, allowing for targeted intervention strategies.
    - Investing in improving customer satisfaction and addressing complaints can significantly reduce churn, as highlighted by the regression analysis.
7. **Resource Allocation:**
    - With high recall models, businesses can identify a broader range of potential churners, allowing for proactive measures to retain customers.
    - Allocate resources effectively by focusing on customers identified by the model as high-risk for churn.
8. **Predictive Maintenance:**
    - Use the model to continuously monitor and predict customer churn, allowing for real-time intervention.
    - Implement a customer feedback loop to adjust and improve the model over time, ensuring it adapts to changing customer behaviors.
9. **Actionable Insights:**
    - Focus on improving customer care scores and resolving complaints efficiently, as these are significant predictors of churn.

- Tailor retention strategies based on customer designation and marital status, which have a considerable impact on churn likelihood.

The **Random Forest** model is the most optimal for predicting customer churn due to its superior performance across multiple metrics. Deploying this model can significantly enhance the life insurance sales business's ability to retain customers by identifying those at high risk of churn and implementing targeted retention strategies. Additionally, continuous monitoring and adaptation of the model will ensure sustained improvement in customer retention efforts.