CMT307 Coursework 1

Implementation And Evaluation Of A Case Study Using Machine Learning Techniques

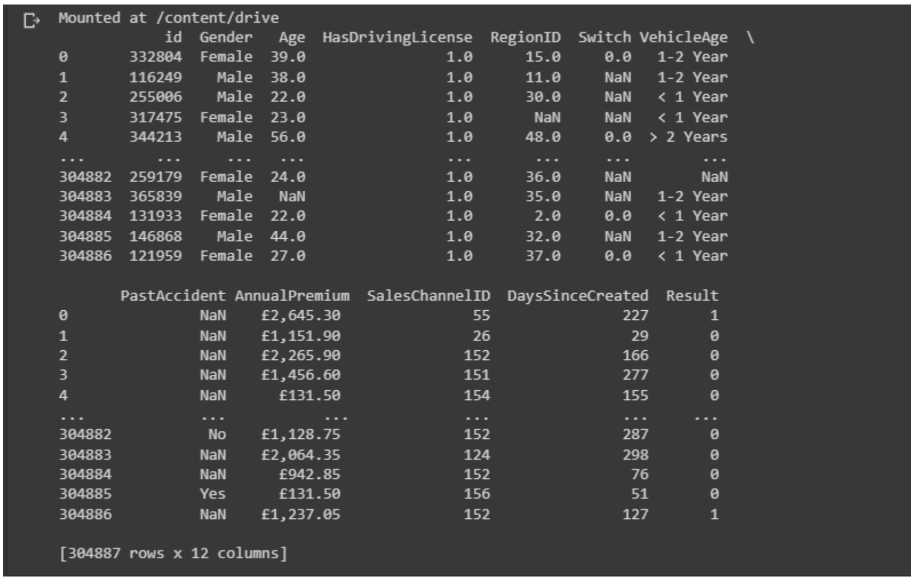
Priyanka Magar 22074213

Table of Contents

Table of Figures

Data Exploration

In machine learning the quality of product is a direct reflection on the data inputted. There are multiple data characteristics that can cause poor results and they must be recognised and addressed so that the results produced are optimum. For example data containing outliers or missing data will lead to unreliable results.



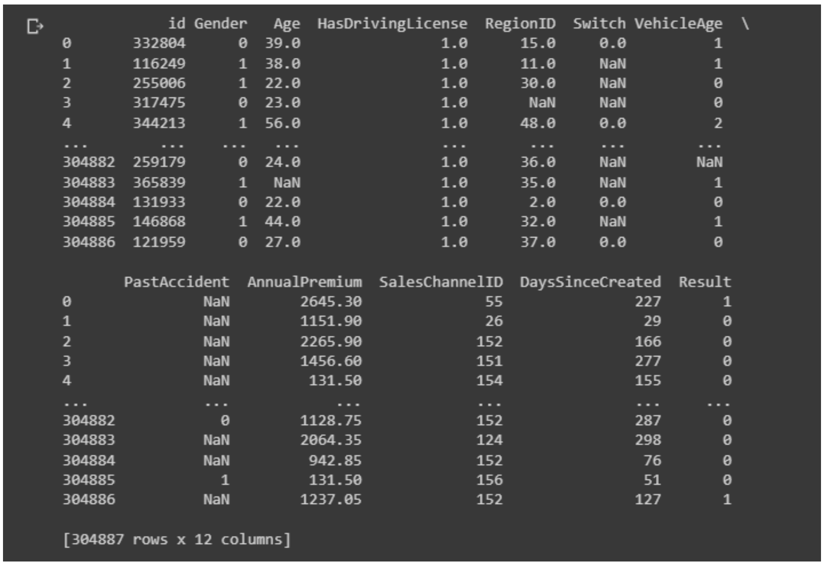
```
Mounted at /content/drive
            id  Gender   Age  HasDrivingLicense  RegionID  Switch VehicleAge  \
0       332804  Female  39.0                1.0      15.0     0.0   1-2 Year
1       116249    Male  38.0                1.0      11.0     NaN   1-2 Year
2       255006    Male  22.0                1.0      30.0     NaN   < 1 Year
3       317475  Female  23.0                1.0       NaN     NaN   < 1 Year
4       344213    Male  56.0                1.0      48.0     0.0  > 2 Years
...        ...     ...   ...                ...       ...     ...        ...
304882  259179  Female  24.0                1.0      36.0     NaN        NaN
304883  365839    Male   NaN                1.0      35.0     NaN   1-2 Year
304884  131933  Female  22.0                1.0       2.0     0.0   < 1 Year
304885  146868    Male  44.0                1.0      32.0     NaN   1-2 Year
304886  121959  Female  27.0                1.0      37.0     0.0   < 1 Year

       PastAccident AnnualPremium  SalesChannelID  DaysSinceCreated  Result
0               NaN     £2,645.30              55               227       1
1               NaN     £1,151.90              26                29       0
2               NaN     £2,265.90             152               166       0
3               NaN     £1,456.60             151               277       0
4               NaN       £131.50             154               155       0
...             ...           ...             ...               ...     ...
304882           No     £1,128.75             152               287       0
304883          NaN     £2,064.35             124               298       0
304884          NaN       £942.85             152                76       0
304885          Yes       £131.50             156                51       0
304886          NaN     £1,237.05             152               127       1

[304887 rows x 12 columns]
```

FIGURE 1: RAW DATA

Initially the raw data was inspected manually. Upon inspection it was noticeable that there were some missing data. Missing data can cause issues especially when using algorithms such as random forest and gradient boosting as they could lead to biased prediction values. The data is not in a uniform format for all the attributes meaning that it must be transformed. The AnnualPremium attribute needed to be transformed to a float value and remove any unnecessary characters. Some of the categorical features also needed to be transformed so that it is more suitable for the models. The gender and PastAccident features were converted to binary and the vehicle age to numerical categories.



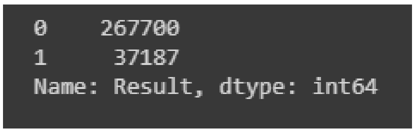
```
              id Gender   Age  HasDrivingLicense  RegionID  Switch VehicleAge  \
0       332804      0  39.0                1.0      15.0     0.0          1
1       116249      1  38.0                1.0      11.0     NaN          1
2       255006      1  22.0                1.0      30.0     NaN          0
3       317475      0  23.0                1.0       NaN     NaN          0
4       344213      1  56.0                1.0      48.0     0.0          2
...        ...    ...   ...                ...       ...     ...        ...
304882  259179      0  24.0                1.0      36.0     NaN        NaN
304883  365839      1   NaN                1.0      35.0     NaN          1
304884  131933      0  22.0                1.0       2.0     0.0          0
304885  146868      1  44.0                1.0      32.0     NaN          1
304886  121959      0  27.0                1.0      37.0     0.0          0

        PastAccident  AnnualPremium  SalesChannelID  DaysSinceCreated  Result
0                NaN        2645.30              55               227       1
1                NaN        1151.90              26                29       0
2                NaN        2265.90             152               166       0
3                NaN        1456.60             151               277       0
4                NaN         131.50             154               155       0
...              ...            ...             ...               ...     ...
304882             0        1128.75             152               287       0
304883           NaN        2064.35             124               298       0
304884           NaN         942.85             152                76       0
304885             1         131.50             156                51       0
304886           NaN        1237.05             152               127       1

[304887 rows x 12 columns]
```

FIGURE 2: DATA AFTER TRANSFORMATION

The data balance for the target variables was checked and it showed that there is a huge imbalance. The major class being the group that are not interested and the minor class being the ones that are interested.



```
0    267700
1     37187
Name: Result, dtype: int64
```

FIGURE 3: RESULT FEATURE IMBALANCE

The data imbalance for all the features were visualized. From the visualisation the only two features that had a big data imbalance was HasDrivingLicence and Result. The result is the target variable which means that its data balance will have a negative effect on the model results due to biased learning. The model might perform well on the majority class but not on the minority class meaning that it has poor generalisation due to not being able to capture the underlying patterns within the minority class. When choosing the evaluation metrics accuracy would not be a good metric as it can be misleading and cause low recall value.

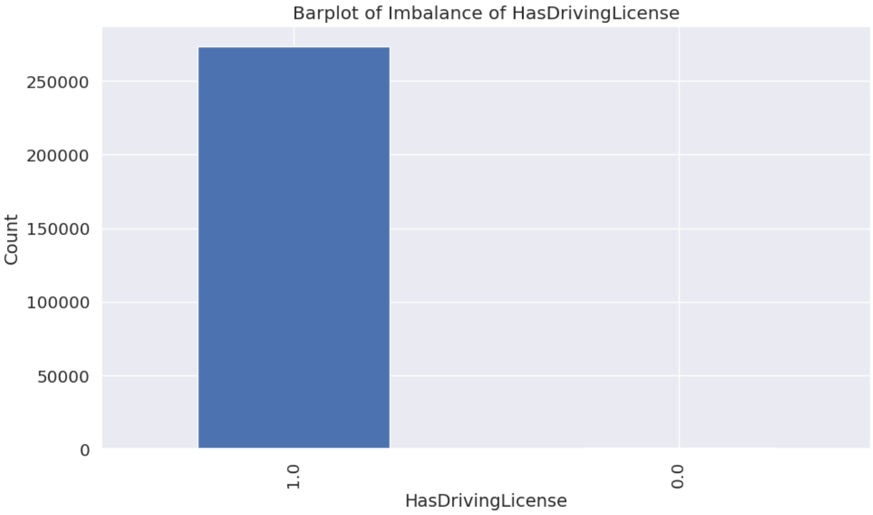


FIGURE 4: HASDRIVINGLICENSE DATA IMBALANCE

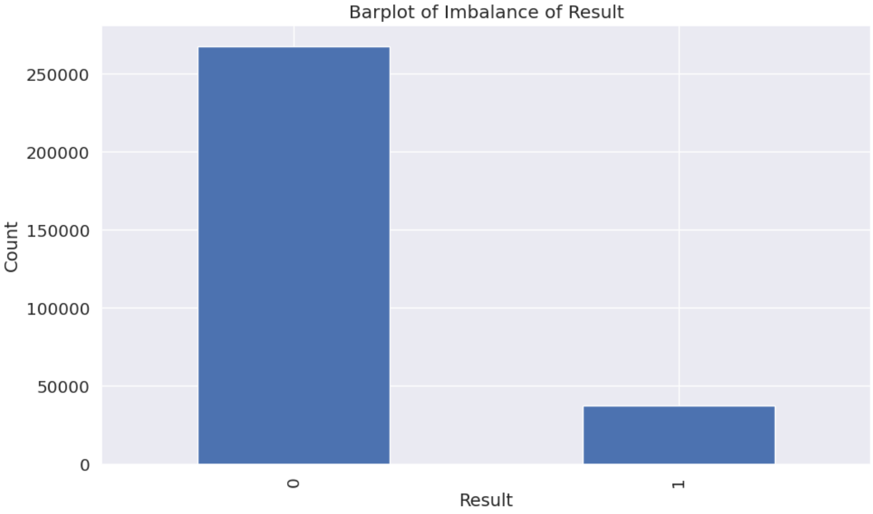
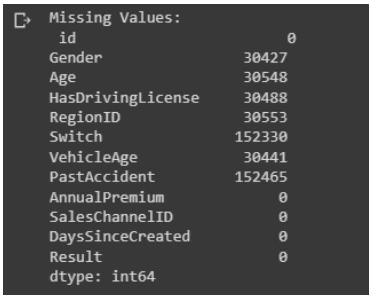


FIGURE 5: RESULT DATA IMBALANCE

The data was also checked for any missing values within the data entries. Missing values can also lead to biased estimates and distort the relationship between the variables and therefore inaccurate results.



```
Missing Values:
 id                     0
Gender              30427
Age                 30548
HasDrivingLicense   30488
RegionID            30553
Switch             152330
VehicleAge          30441
PastAccident       152465
AnnualPremium           0
SalesChannelID          0
DaysSinceCreated        0
Result                  0
dtype: int64
```

FIGURE 6: MISSING VALUES

Each numerical feature was then further studied to understand its distribution to deal with the missing values.

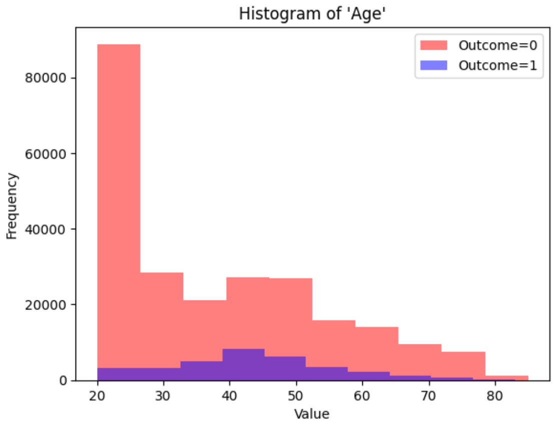


Figure 7: Age histogram

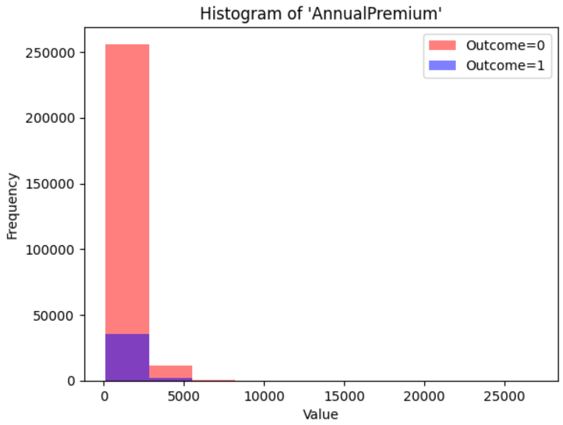


Figure 8: Annual premium histogram

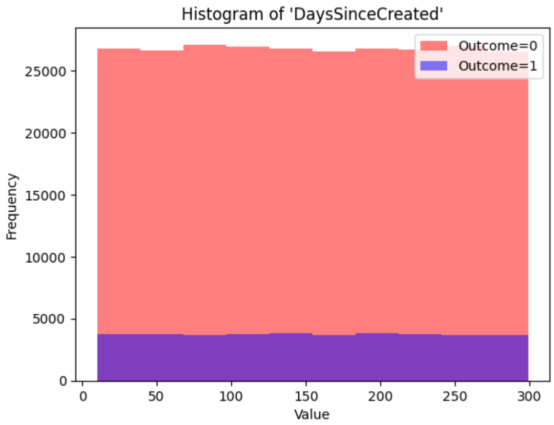


Figure 9: Days since created Histogram

The extreme values of the outliers can cause various issues depending on the implementation of the models. For example in a decision tree model the structure of the model can be disrupted causing noise that outshine the meaningful relationships. To visualise the outliers box plots were created.

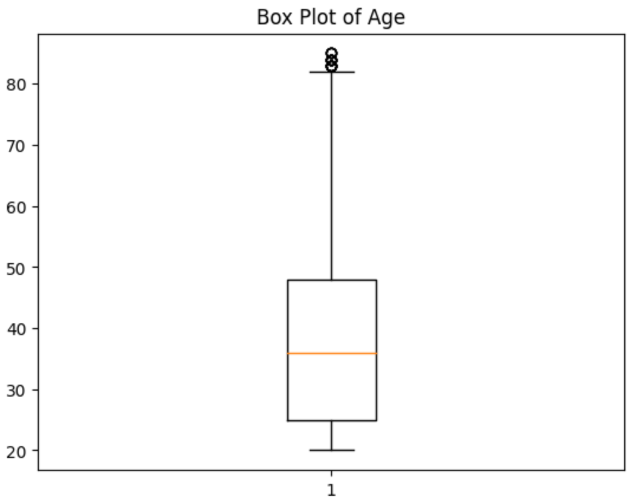


FIGURE 10: AGE BOXPLOT

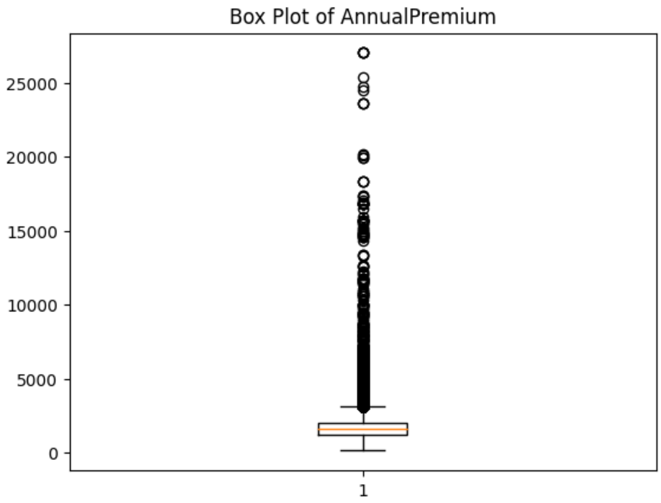


FIGURE 11: ANNUALPREMIUM BOXPLOT

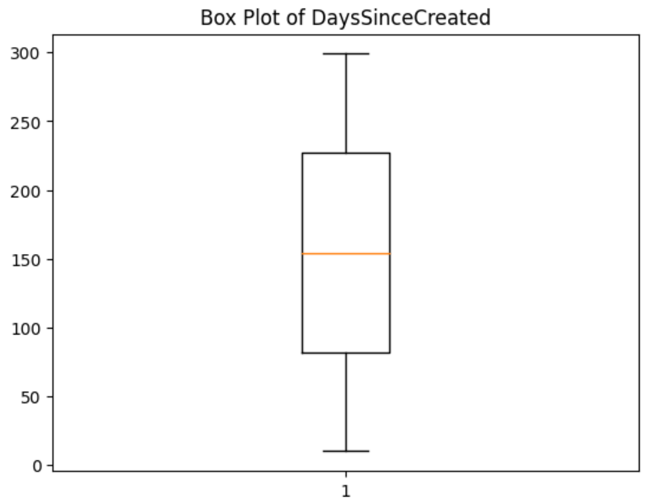


FIGURE 12: DAYSSINCECREATED BOXPLOT

Including unnecessary features that do not have any correlation to the target variable will cause unnecessary noise that could overshadow the meaningful relationships. To understand the relationships between each x variable features and the target variables a heatmap of the correlations was created. The categorical features were one hot encoded.

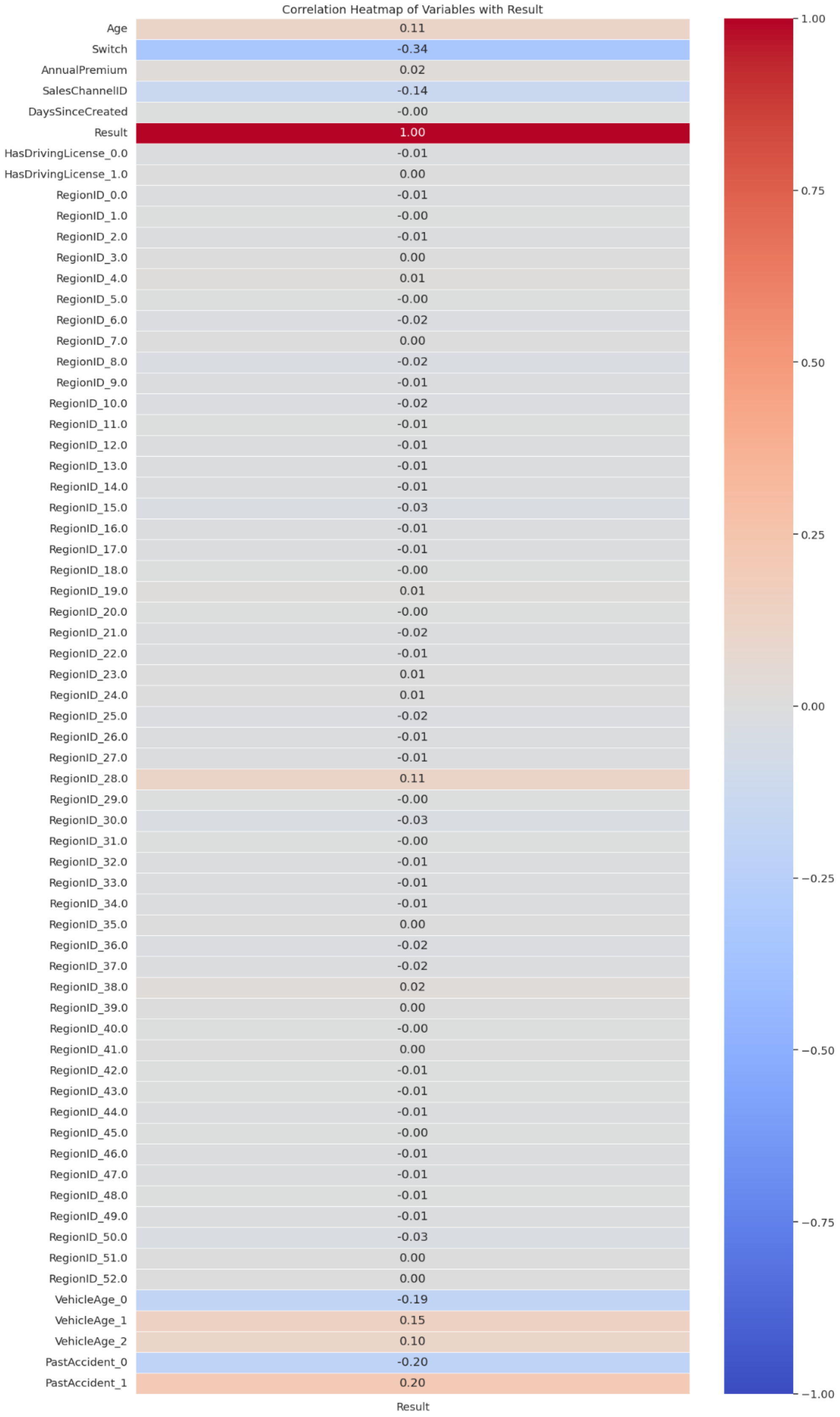


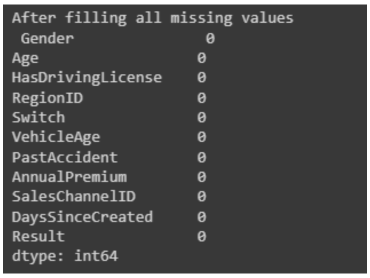
FIGURE 13: HEATMAP OF CORRELATION TO TARGET VARIABLE

Data Preprocessing

From the initial exploration there were multiple characteristics of the data that needed to be improved to provide quality results. While exploring the data some preprocessing techniques had already been implemented such as formatting the data.

Missing Data

There are multiple methods of dealing with missing data to minimise the negative effects. The missing data in the HasDrivingLicence feature seems to be an entry error as there are no entries for the class that do not have driving licence. In real world situations it can be argued that even though an individual might not have a driving licence they could still be interested in car insurance such as for their kids and family members. In this case every missing entry was then replaced by 0 indicating that they did not have a driving licence. The age when plotted in the histogram has a normal distribution meaning that the missing values were replaced by the median. For the other categorical features a new category was created to deal indicate that it has missing values. Alternatively the entries with missing values can also be deleted for experimental purposes, however this would greatly affect the sample size of the data due to the large quantity of missing values.



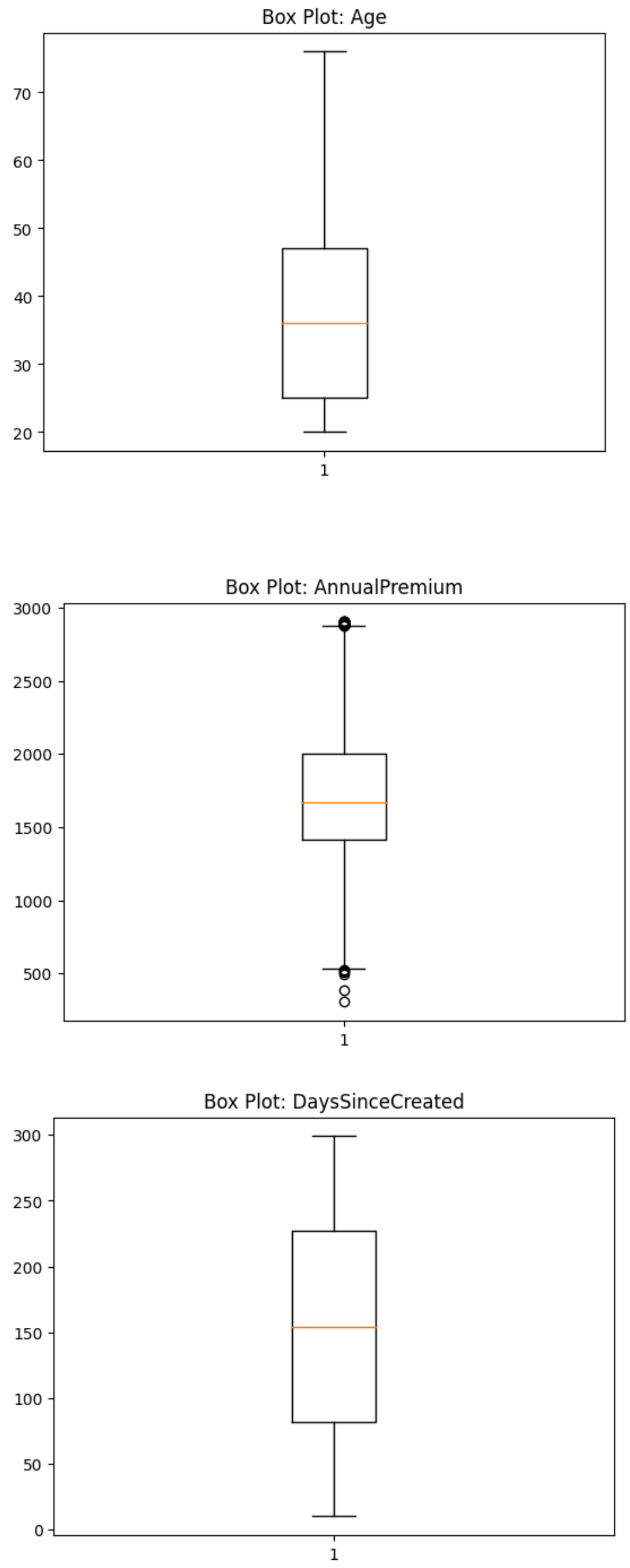
```
After filling all missing values
 Gender               0
Age                  0
HasDrivingLicense    0
RegionID             0
Switch               0
VehicleAge           0
PastAccident         0
AnnualPremium        0
SalesChannelID       0
DaysSinceCreated     0
Result               0
dtype: int64
```

FIGURE 14: AFTER DEALING WITH MISSING VALUES

Outliers

From the box plot it can be seen that there is a large number of outliers for the AnnualPremium feature and a small number of outliers for the age feature. Initially these data entries will be removed, however in real life situations these values may be outliers but would not necessarily be wrongful entries. The highest value for age is about 90 years old which is a valid value. The outliers in the annual premiums are also plausible as majority of the population would likely have a premium within the mean value, however there might be individuals that have very expensive cars that have a higher rate premiums. Initially the outliers are removed to start the implementation of the models, however it might be a good idea to keep them depending on the result of the models.

To deal with the outliers any values that are outside of the 1.25 range were removed.



Box Plot: Age

Box Plot: AnnualPremium

Box Plot: DaysSinceCreated

FIGURE 15: BOX PLOTS AFTER REMOVING OUTLIERS

Feature Engineering

Initially the features with very low correlations were removed from the data. DaysSinceCreated has a correlation of 0 meaning that it was removed. Annual premium, Has driving licence and RegionID have a very low correlation so that was also removed. Entries with RegionID 28 had a higher correlation of 0.11 so this means that during the experimental phase it might be worth including this feature. Removing these features decreases the issues that may be caused by high dimensions and reducing noise.

Dealing with data Imbalance

There are three main method of dealing with imbalance. The data can be oversampled which means that we could add more entries to the dataset, under sampling where you randomly remove the entries from the majority class to balance the data and leaving the data imbalanced. In the case of this study oversampling is not possible, so the other two options are viable to experiment with. To begin with the data was split into majority and minority class so that the data imbalance can be dealt with by under sampling the data. The sampled data is also shuffled to remove any biases caused by the order. Leaving the data unbalanced can be something that can be tested during the experimental phase.

Model Implementation

The algorithms utilised are determined by the data that is available. The data contains both categorical and numerical features so an algorithm that can process both of these types of data must be chosen. If possible, the data types could also be transformed if necessary. The data available is labelled meaning that supervised learning is going to be the most appropriate choice. Further choices can be made depending on experimentation.

The models chosen for this case are XGBoost, random forest and logistic regression. XGBoost is a gradient boosting algorithm that can handle complex relationships in a dataset and is known for its high performance. Gradient boosting models combine multiple weak learners into a strong ensemble, reducing both bias and variance.

Logistic regression was chosen because it aligns with the insurance providers goal of understanding the impact of various features on how likely they are to be interested in car insurance. It is an efficient algorithm and serves as a good baseline model for understanding the relationships between the features and the target variable. This model however assumes linear relationship.

The final model chosen is random forest. Random forest has an ability to capture non-linear relationships and capturing hidden patterns, however requires a much larger computational power.

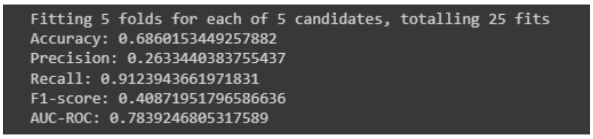
The models have all used techniques to increase the quality of the results. The models all make use of the randomizedSearchCV functions that takes predefined parameters and tests various combinations for hyperparameter tuning with 5 fold cross validation. Tuning the hyperparameters optimises the model with the goal of finding the best performing models. Tuning the hyperparameters in conjunction with having 5 fold cross validation assesses the models performance on new and unseen data and helps deal with issues regarding overfitting and underfitting. This is accomplished by splitting the available data into multiple subsets and evaluating the performance iteratively. This ensures that the model is much more robust. Once the best model has been found the models are further optimised by using the bagging technique. It resamples the data into subsets decreasing error by decreasing the variance in the results caused by unstable learners.

Ensemble

Ensemble model is created by combining various models to improve the performance. The performance of an ensemble model is much higher than a singular model as the diversity makes it more stable. It becomes less susceptible to biases, variance and overfitting, however due to having multiple parts it becomes more susceptible to unpredictability as one error can have a big effect on the predictions.

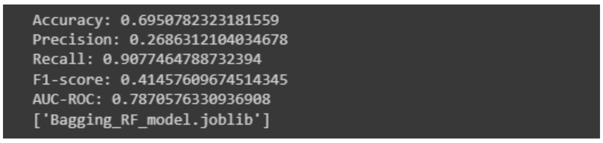
Results

The five evaluations metrics used are accuracy, recall, precision, f1 score and AUC. It was previously determined that the data available was not balanced meaning that accuracy might not be the best metric to use when evaluating the performance of the models, there it is calculated just for the purpose of relative comparison of the 3 models. AUC value shows the models ability to differentiate between different classes across different probability thresholds.



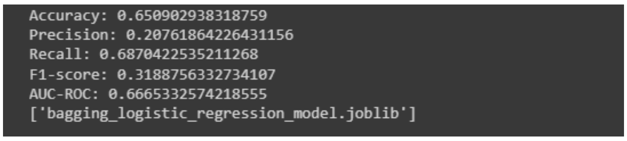
```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Accuracy: 0.6860153449257882
Precision: 0.2633440383755437
Recall: 0.9123943661971831
F1-score: 0.40871951796586636
AUC-ROC: 0.7839246805317589
```

FIGURE 16: XGBOOST RESULTS



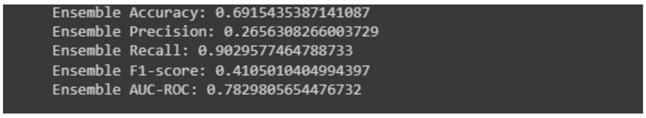
```
Accuracy: 0.6950782323181559
Precision: 0.2686312104034678
Recall: 0.9077464788732394
F1-score: 0.41457609674514345
AUC-ROC: 0.7870576330936908
['Bagging_RF_model.joblib']
```

FIGURE 17: RANDOM FOREST RESULTS



```
Accuracy: 0.650902938318759
Precision: 0.20761864226431156
Recall: 0.6870422535211268
F1-score: 0.3188756332734107
AUC-ROC: 0.6665332574218555
['bagging_logistic_regression_model.joblib']
```

FIGURE 18: LOGISTIC REGRESSION RESULTS



```
Ensemble Accuracy: 0.6915435387141087
Ensemble Precision: 0.2656308266003729
Ensemble Recall: 0.9029577464788733
Ensemble F1-score: 0.4105010404994397
Ensemble AUC-ROC: 0.7829805654476732
```

FIGURE 19: ENSEMBLE MODEL RESULTS

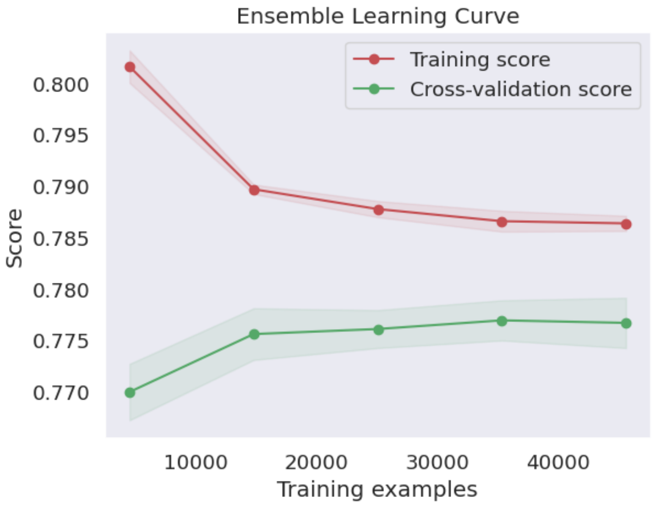


FIGURE 20: LEARNING CURVE

Evaluation

The results from all three model as well as the ensemble model shows that there is a high recall score, but a low precision score leading to a f1 score in the middle. In this case this indicates the models are very good at identifying customers that are interested in purchasing car insurance and therefore less likely to miss out customers who are interested in car insurance. The precision is the model's positive prediction. It is the accuracy of the entries that were labelled as being interested in car insurance. The low precision indicates that the model labels high number of people who would not be interested in car insurance as being interested. This means that if the company is looking for potential car insurance buyers, then they should reach out to all customers that are labelled as being interested, however they should expect that some of those customers might reply by saying that they are not interested. The models ensure that it maximises the reach of the customers that are genuinely interested, however the company will be wasting resources by reaching out to customers who may not be interested. There is a risk that reaching out to customers that are not interested can negatively affect their relationship by causing annoyance.

The learning curve of the training shows that the model is learning very well from the data and the decrease in training accuracy shows that it is likely to be less susceptible to overfitting and going to be generalised. The convergence of cross validation shows that there is a good balance between bias and variance. The gap between the plots suggests that the model may be a little overfitted but this is expected due to the data imbalance.

The results created show that the model provides a good starting point for the company and that potential customers can be reached however the models could be further optimised so that the negative drawbacks can be minimised.

Future Work

There are many aspects of the study that was not able to be completed due to the lack of resources. The two main aspects that were lacking were in regard to time constrains and the computational power available. In the future the model can be further optimized firstly by improving the data preprocessing. Although multiple preprocessing methods have been used it was not determined whether the techniques improved or worsened the results. One example of this is dealing with missing values. If the missing values were not filled in, due to the nature of the algorithms and their ability to handle missing data the results could have been better. Trying different interquartile range when removing outliers could have also improved the results. Furthermore, testing out other algorithms that singularly create better results such as testing the SVM algorithm could have also improved the model. This was not possible in this scope of the study due to its high demand in computational power.

Bibliography

Feature Engineering for Machine Learning and Data Analytics (2020). S.l.: CRC PRESS.

González, S. et al. (2020) ‘A practical tutorial on bagging and boosting based ensembles for Machine Learning: Algorithms, software tools, performance study, Practical Perspectives and Opportunities’, Information Fusion, 64, pp. 205–237. doi:10.1016/j.inffus.2020.07.007.

Gopagoni, D.R., Lakshmi, P.V. and Siripurapu, P. (2020) ‘Predicting the sales conversion rate of car insurance promotional calls’, Rising Threats in Expert Applications and Solutions, pp. 321–329. doi:10.1007/978-981-15-6014-9_37.

Hristeva, T. (2022) ‘Application of graphic processing units in deep learning algorithms’, 17TH INTERNATIONAL CONFERENCE ON CONCENTRATOR PHOTOVOLTAIC SYSTEMS (CPV-17) [Preprint]. doi:10.1063/5.0091076.

Kaur, H., Pannu, H.S. and Malhi, A.K. (2019) ‘A systematic review on imbalanced data challenges in machine learning’, ACM Computing Surveys, 52(4), pp. 1–36. doi:10.1145/3343440.

Mare, C. et al. (2022) ‘Machine learning models for predicting Romanian farmers’ purchase of Crop Insurance’, Mathematics, 10(19), p. 3625. doi:10.3390/math10193625.