MACHINE LEARNING SOLUTION DEVELOPMENT REPORT

PREPARED FOR

Travel Insurance LTD

PREPARED BY

Data Science Solutions

Project Report

1. The Classification Learning Problem

Goal

The goal of working on this classification learning problem was to build a machine learning model that could, to a high degree of accuracy, predict whether a customer would make an insurance claim.

Methodology

The first stage of the experiment was to observe the Training dataset we had at hand to build this model. (see Figure 1)

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	Class
0	4.92	-13.23	330	13.22	-5.51	41.10	16.72	-56.98	-748.72	20	9.75	259.66	8.00	-0.35	4.76	False
1	-9.33	-27.72	3	12.78	-9.12	5.31	9.90	-30.98	-436.72	2	-1.56	131.66	0.76	1.79	NaN	False
2	-15.09	-26.28	6	13.54	-7.75	5.67	9.93	-34.98	-482.72	2	-2.58	63.66	0.64	2.02	NaN	True
3	-18.09	-24.60	30	13.28	-8.61	3.51	10.02	-30.98	-414.72	2	-7.08	133.66	0.57	1.98	NaN	True
4	6.12	-8.64	300	14.94	-1.97	29.40	19.32	-36.98	-428.72	20	9.09	229.66	6.70	-3.83	NaN	True
5	6.72	-12.90	325	11.54	-3.94	16.50	14.82	-76.98	-408.72	20	14.55	39.66	9.25	-3.27	7.02	True
6	-8.61	-23.79	44	12.48	-7.29	0.96	10.24	-34.98	-468.72	2	-0.36	95.66	0.56	1.98	6.53	True
7	5.85	-9.39	390	13.68	-2.43	34.35	18.27	-36.98	-588.72	20	13.56	109.66	5.50	-3.84	NaN	False
8	-13.80	-27.00	25	11.72	-9.73	5.58	9.47	-34.98	-450.72	2	-0.27	83.66	0.56	4.96	NaN	False
9	4.02	-11.46	95	10.38	-6.33	30.60	17.22	-56.98	-758.72	20	12.36	139.66	5.70	-2.87	NaN	True

Figure 1 (DataFrame of the Classification Learning Problem)

One issue that was immediately noticed with this dataset was that the feature F15 had missing values. This is not ideal when building a machine learning model as the feature with missing values could be an important determining feature of the class of a customer (Witten. 2017. 61-62). To resolve this problem in the dataset without discarding many rows worth of data and reducing the training data, was to find the mean of the values that were present in F15 and replace the missing values in that column with the mean value calculated (see Figure 2).

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	4.92	-13.23	330.0	13.22	-5.51	41.10	16.72	-56.98	-748.72	20.0	9.75	259.66	8.00	-0.35	4.76000
1	-9.33	-27.72	3.0	12.78	-9.12	5.31	9.90	-30.98	-436.72	2.0	-1.56	131.66	0.76	1.79	5.75531
2	-15.09	-26.28	6.0	13.54	-7.75	5.67	9.93	-34.98	-482.72	2.0	-2.58	63.66	0.64	2.02	5.75531
3	-18.09	-24.60	30.0	13.28	-8.61	3.51	10.02	-30.98	-414.72	2.0	-7.08	133.66	0.57	1.98	5.75531
4	6.12	-8.64	300.0	14.94	-1.97	29.40	19.32	-36.98	-428.72	20.0	9.09	229.66	6.70	-3.83	5.75531
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	-1.35	-20.76	40.0	11.58	-8.75	41.85	14.52	-56.98	-828.72	20.0	1.08	159.66	5.90	3.54	5.75531
1496	-12.90	-25.80	8.0	12.38	-9.51	0.51	9.97	-34.98	-424.72	2.0	-6.21	75.66	0.54	3.84	5.75531
1497	-3.78	-14.97	155.0	9.20	-4.18	40.05	15.17	-56.98	-748.72	20.0	6.96	-50.34	6.15	-0.76	4.42000
1498	1.38	-14.97	190.0	13.36	-6.31	33.15	14.47	-36.98	-628.72	20.0	10.20	269.66	8.30	-2.63	3.29000
1499	8.16	-7.86	330.0	16.14	-3.18	18.00	19.12	-36.98	-478.72	20.0	14.10	219.66	7.00	-3.54	5.75531

1500 rows × 15 columns

Figure 2 (Dataframe where missing F15 values (index 14) are replaced with the mean values of the column)

The dataset was then visualized and correlated with all the variables being compared to one another to check for any false predictors in the dataset (see Figure 3).

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
-0.079108	-0.057144	-0.152249	-0.024064	-0.068795	-0.191255	-0.080522	0.144057	0.266599	-0.203649	-0.058308	-0.050738	-0.135190	0.079035	0.729540

Figure 3 (Correlation values of each value with the Class attribute)

The Feature columns and Class column were converted into NumPy arrays, separated and split into their respective training and testing counterparts using the method of Stratified K-fold Cross-Validation. An advantage of using Stratified K-fold Cross-validation is the dataset is split into a predetermined amount training and testing sets, which means the model trained will more likely have less bias than the standard train/test split method. The K value selected when constructing this classification model was K = 10, as this value is deemed to be an optimal value in terms of the tradeoff of computation, variance and bias for a reasonably sized dataset (Kohav. 1995). Particularly an advantage Stratified K-fold Cross-Validation has over traditional K-fold Cross validation is the results of Stratified K-fold tends to have slightly less bias and variance over traditional K-fold in classification models (Kohav. 1995). The Features data was then normalised using Standard Scalar.

Once the data was pre-processed, the Decision Tree Classifier model, Logistic Regression model, KNN model and a further SVM model were constructed, with the best model being used to predict the values on the unseen data once it had been pre-processed.

Discussion and Evaluation of the Models

All of the model's Cross-Validation accuracy scores can be observed below (see Figure 4). The model with the highest Cross-Validation accuracy score out of all the classification models built on this dataset was the SVM model with the linear kernel. Numerous other SVM kernels were trialled in the model building process (Polynomial, Gaussian, Sigmoid, RBF), although the linear kernel noticeably produced a model with much higher classification accuracy. Possible explanations for why this specific model has a high accuracy is likely due to the linear separability of the Class attribute. This would also explain why the logistic regression model performed well, while models more suited to finding more complex relationships with non-linearly separable data generally performed worse (KNN and Decision Tree Classifier models). This result remains true even with the optimized parameters of the KNN model, with the best K-value being found as used using Grid Search (Geron. 2019. 76-78). Different parameters of the Decision Tree were also tried with entropy loss and the experimentation of the maximum number of branches, the best performing parameters being the use of Gini impurity along with the maximum number of branches (4 branches being the smallest number of branches possible with the highest accuracy score, specifically selected in order to reduce potential overfitting). Despite similar accuracy scores, the SVM model is less likely to overfit the data in comparison to the logistic regression (Chang. 2020). Lastly, SVMs in particular work well with the Standard Scaler method the data was standardized with (Geron. 2019. 152-154).

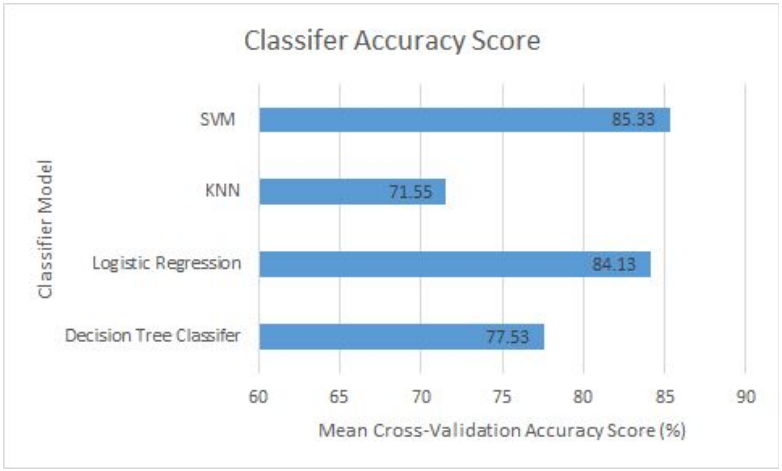

Figure 4 (Classifier Accuracy Score of all trialled models)

Conclusion

In conclusion, the SVM model with a linear kernel is the most superior choice of model trialled for this classification learning problem. The success of the SVM model, as well as the logistic regression model, suggest to us the Class attribute in whether a customer will likely make an insurance claim or not is linearly separable. In future investigations where more optimization of prediction accuracy is necessary, a larger amount of training data would be beneficial and welcome. A further investigation into parameter selection of the current models as well as a development neural network models may also lead to an improvement in prediction accuracy.

2. The Regression Learning Problem

Goal

The goal of working on this regression learning problem was to build a machine learning model that could to a high degree of accuracy predict how much a customer's claim would be worth.

Methodology

Similarly to the classification problem, the first step taken was to observe the dataset we were given to build the model (see Figure 5).

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	Target
0	16.56	12.42	-236.06	Rest	-98.88	529.56	4.54	379.54	1	1	7.30	High	-15085.87	-12.93	-39.42	1734.58	3616.82
1	11.72	12.46	-190.06	Rest	-59.22	493.11	0.05	402.78	5	3	-1.28	Very low	-15782.44	-8.55	-35.61	1672.70	3342.88
2	4.34	2.74	-201.20	UK	-228.48	563.79	1.22	147.35	4	4	8.28	Low	-10526.01	-9.66	-29.10	1462.86	0.00
3	12.76	2.58	-282.26	UK	-173.28	536.94	0.25	113.49	4	3	6.26	Low	-8327.14	-19.23	-34.59	809.46	1742.65
4	11.10	9.82	-242.86	USA	-193.14	617.52	9.15	343.64	8	6	-6.88	Very low	-14434.13	-9.45	-46.14	1435.90	373.56
5	6.36	4.48	-151.96	Rest	-331.80	1043.19	63.10	-170.49	7	5	12.68	Very low	-13858.77	-15.18	-36.30	2384.08	0.00
6	5.30	1.00	-189.06	USA	-336.33	472.56	2.19	290.55	9	3	-0.32	Very high	-17261.41	-15.00	-32.88	1384.70	173.96
7	4.18	10.96	-207.24	Europe	-198.81	568.05	38.18	-202.04	5	2	0.52	High	-10742.10	-9.66	-41.52	1669.36	1661.04
8	10.98	5.68	-241.58	UK	-121.62	498.21	8.78	-152.50	2	2	0.78	High	-13952.70	-3.69	-27.84	2268.44	2135.45
9	11.76	6.90	-156.24	USA	-201.60	574.59	0.95	88.56	7	4	1.42	High	-30289.83	-13.86	-34.77	820.86	1088.68

Figure 5 (DataFrame of the Regression Learning Problem)

The data was then visualized and correlated to check for false predictors (Figure 6).

F1	F2	F3	F5	F6	F7	F8	F9	F10	F11	F13	F14	F15	F16
0.346373	0.357137	-0.015412	0.430990	0.031407	0.027713	-0.300890	0.193704	-0.236200	0.011330	-0.029194	-0.020443	-0.018677	-0.263022

Figure 6 (Correlation values of each value with the Target attribute)

What we can see in Figure 6 is that a few of the features have values that are recorded categorically and not continuously. To successfully build our models, there had to change and code those categorical variables as dummy variables (see Figure 7).

F11	F13	F14	F15	F16	Target	F4_Europe	F4_Rest	F4_UK	F4_USA	F12_High	F12_Low	F12_Medium	F12_Very high	F12_Very low
7.30	-15085.87	-12.93	-39.42	1734.58	3616.82	0	1	0	0	1	0	0	0	0
-1.28	-15782.44	-8.55	-35.61	1672.70	3342.88	0	1	0	0	0	0	0	0	1
8.28	-10526.01	-9.66	-29.10	1462.86	0.00	0	0	1	0	0	1	0	0	0
6.26	-8327.14	-19.23	-34.59	809.46	1742.65	0	0	1	0	0	1	0	0	0
-6.88	-14434.13	-9.45	-46.14	1435.90	373.56	0	0	0	1	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12.32	-16977.67	5.01	-29.67	1420.16	1515.56	0	1	0	0	0	0	0	1	0
7.40	-13927.89	0.66	-36.06	1321.82	1528.48	0	0	1	0	0	0	1	0	0
7.02	-18373.26	-7.68	-39.42	1905.86	323.67	1	0	0	0	0	0	1	0	0
1.68	-22863.72	-3.12	-54.36	2424.42	109.19	0	0	0	1	0	0	0	1	0
-7.66	-4070.71	-0.36	-46.98	2222.52	0.00	0	0	1	0	0	0	0	0	1

Figure 7 (partial DataFrame displaying the dummy variables encoded)

Once this was completed we used standard K-fold Cross-Validation (K=10) to split our data into training and testing groups as well as normalising the data using Standard Scalar.

As many claims were valued to be zero in the DataFrame, rather than just running regression algorithm on the whole training dataset it was decided to increase the overall accuracy of the predictions to run a logistic regression classification algorithm to predict the customer claims that were likely to be zero and what claims were likely to be a non-zero value.

Another copy of the dataset was then generated with all the zero values rows from the Target column removed and the dataset was split and normalised using MinMaxScaler. The three regression models trialled were Multiple Linear Regression, Decision Tree Regressor and Random Forest Tree Regressor. Then the logistic regression model and most successful regression model were selected and applied to predict the target values on the unseen data once it had been pre-processed.

Discussion and Evaluation of the Models

Classification Model (Logistic Regression)

The reason why the logistic regression algorithm was built on the dataset before running the regression models was to precisely classify what claims were likely to be valued as zero. Building the regression models straight away provided many predictions for the zero values that were minus numbers, which provides us with counterproductive predictions given this specific task. In order to mitigate this problem and be more precise in our overall prediction, a logistic regression algorithm was trialled. The accuracy observed by using the classification algorithm to predict the zero and non-zero values was around 98% (see Figure 8).

```
Logistic Regression Accuracy: 0.9860
```

Figure 8 (Logistic Regression Cross Validation Accuracy)

Showing that the model can predict the values of claims that are zero and non-zero values to a very high degree of accuracy suggests classifying the zero and non-zero values is an easy learning task (less concern of overfitting even with high accuracy metric) in which the classes are linearly separable. Using logistic regression prior to using the regression algorithms also produce a noticeable reduction in the number of predictions with a minus value, it was decided that it is necessary to supplement the regression models by running this classification algorithm before executing the regression model to predict the zero and non-zero values.

Regression Models

The r2 score was the selected metric to evaluate the models in this investigation as it measures the variance that can be explained by a model. The r2 Cross-Validation score of all the regression models trialled can be viewed below (see Figure 9). The model that produced the highest r2 score was the Multiple Linear regression model followed by the Random Forest Regressor and then the Decision Tree Classifier, which performed the worst by quite a substantial margin according to this specific evaluation metric. Due to the Multiple Linear Regression model performing so well, we could assume there is a relatively linear relationship between the features and the Target variable. A reason why the Random Forest Regressor likely outperformed the Decision Tree Regressor is likely due to Random Forest being a more superior model compared to the Decision Tree due to it being an ensemble of Decision Trees in which the value is predicted is obtained through multiple different Decision Tree averages (Géron. 2019. 197-200). Both models likely failed to perform as well as the Multiple Linear Regression model due to the possible linear relationship of the features and Target attributes. A further reason is that the training dataset needs to be larger for the models to be more accurate, as due to running the classification model prior, the training size for the regression model was reduced by approximately ⅓.

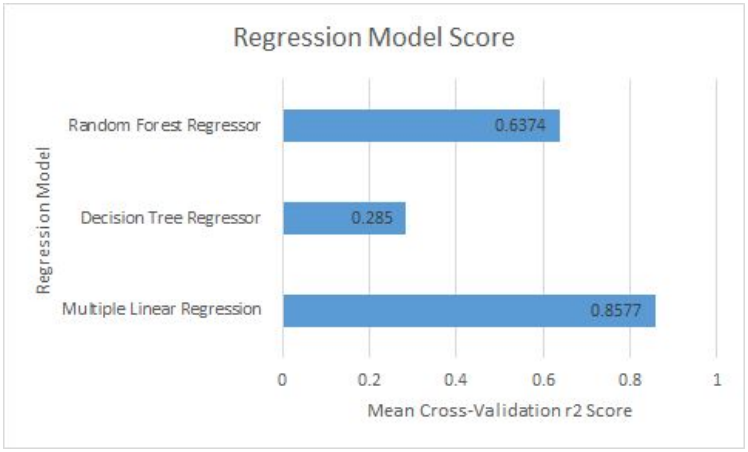

Figure 9 (Regression Model Score of all trialled models)

Conclusion

Overall, out of all the models trialled for this regression learning problem, the multiple linear regression algorithm provided the best predictions on the testing sets according to the r2 evaluation metric used. This suggests there is a linear relationship between the features and the Target attribute. While multiple linear regression can be used, the other regression models provided a disappointing result by comparison. Potential solutions for increasing accuracy and building better models include expanding the training data. Alternatively, furthering trialling of different types of regression models including artificial neural networks may provide better prediction accuracy. With this all in mind, it is currently recommended to use the combination of a logistic regression classifier along with multiple linear regression as prescribed in this report to predict the value of a customer's claim.

References

Chang, Anthony C. 2020. Intelligence-Based Medicine: Artificial Intelligence and Human Cognition in Clinical Medicine and Healthcare. London, United Kingdom: Academic Press is an imprint of Elsevier.

Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts,Tools, and Techniques to Build Intelligent Systems. Sebastopol: O'Reilly Media, Incorporated, 2019. Accessed January 17, 2019. ProQuest Ebook Central.

Kohavi , Ron. 1995. “A Study of Cross-Validation and Bootstrap for Accuracy ...” Accessed January 15. https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection.

Witten, I. H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2017. Data Mining : Practical Machine Learning Tools and Techniques. Vol. Fourth edition. Cambridge, MA: Morgan Kaufmann. http://0-search.ebscohost.com.serlib0.essex.ac.uk/login.aspx?direct=true&db=nlebk&AN=1214611&site=ehost-live.