Business Presentation

Contents

Renewable energy sources play an increasingly important role in the global energy mix, as the effort to reduce the environmental impact of energy production increases.

Out of all the renewable energy alternatives, wind energy is one of the most developed technologies worldwide. The U.S Department of Energy has put together a guide to achieving operational efficiency using predictive maintenance practices.

Predictive maintenance uses sensor information and analysis methods to measure and predict degradation and future component capability. The idea behind predictive maintenance is that failure patterns are predictable and if component failure can be predicted accurately and the component is replaced before it fails, the costs of operation and maintenance will be much lower.

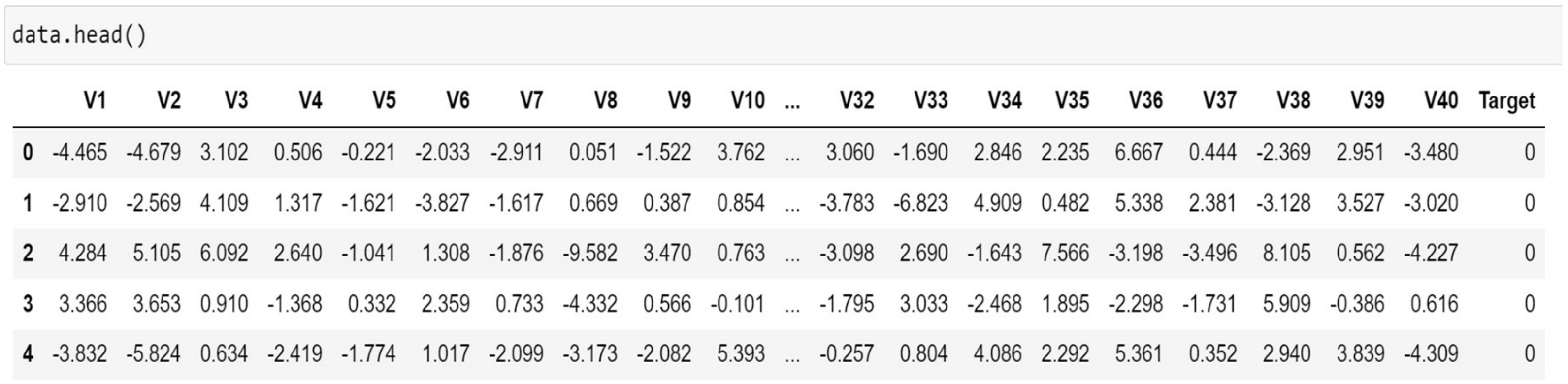
Business Problem Overview and Solution Approach

“ReneWind” is a company working on improving the machinery/processes involved in the production of wind energy using machine learning and has collected data of generator failure of wind turbines using sensors.

The objective is to build various classification models, tune them and find the best one that will help identify failures so that the generator could be repaired before failing/breaking and the overall maintenance cost of the generators can be brought down.

Data Overview

- The train data has 41 columns and 40000 rows
- the data collected through sensors is confidential, so we do not have real name of them
- All columns are numerical
- missing values in the data observed. I replaced them by median.
- There are some outliers in the data
- “1” in the target variables should be considered as “failure” and “0” will represent “No failure”.

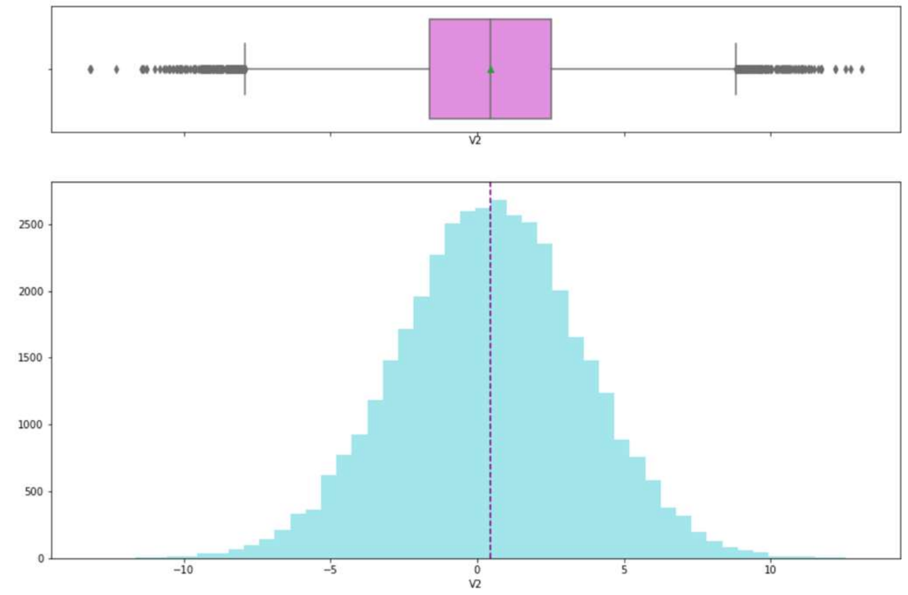


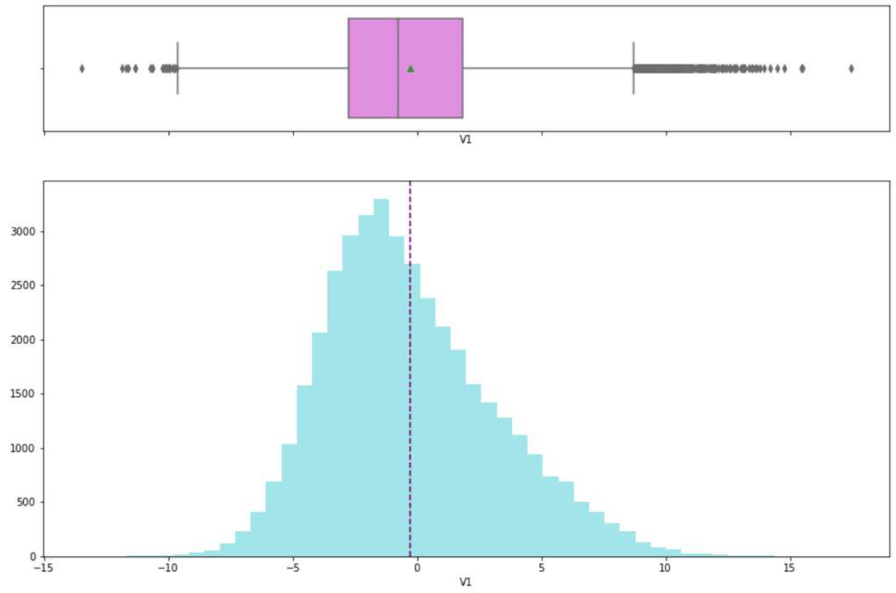
```
data.head()
```

```
       V1      V2     V3      V4      V5      V6      V7      V8      V9     V10  ...     V32     V33     V34    V35     V36     V37     V38     V39     V40  Target
0  -4.465  -4.679  3.102   0.506  -0.221  -2.033  -2.911   0.051  -1.522   3.762  ...   3.060  -1.690   2.846  2.235   6.667   0.444  -2.369   2.951  -3.480       0
1  -2.910  -2.569  4.109   1.317  -1.621  -3.827  -1.617   0.669   0.387   0.854  ...  -3.783  -6.823   4.909  0.482   5.338   2.381  -3.128   3.527  -3.020       0
2   4.284   5.105  6.092   2.640  -1.041   1.308  -1.876  -9.582   3.470   0.763  ...  -3.098   2.690  -1.643  7.566  -3.198  -3.496   8.105   0.562  -4.227       0
3   3.366   3.653  0.910  -1.368   0.332   2.359   0.733  -4.332   0.566  -0.101  ...  -1.795   3.033  -2.468  1.895  -2.298  -1.731   5.909  -0.386   0.616       0
4  -3.832  -5.824  0.634  -2.419  -1.774   1.017  -2.099  -3.173  -2.082   5.393  ...  -0.257   0.804   4.086  2.292   5.361   0.352   2.940   3.839  -4.309       0
```

EDA

- All the columns have a normal distribution. there are two examples of them in the below plots.
- Also from the box plot, we can see that there are a few outliers in most of them which I dealt with them.





Model Performance Summary

- since we have two separate data frames for this project [train data and test data], we do not need to split the data to train and test
- but it is good to split data (train and validation) to see which model performance is better

- independent variables
- x = all my data but: Target
- dependent variable
- y = Target

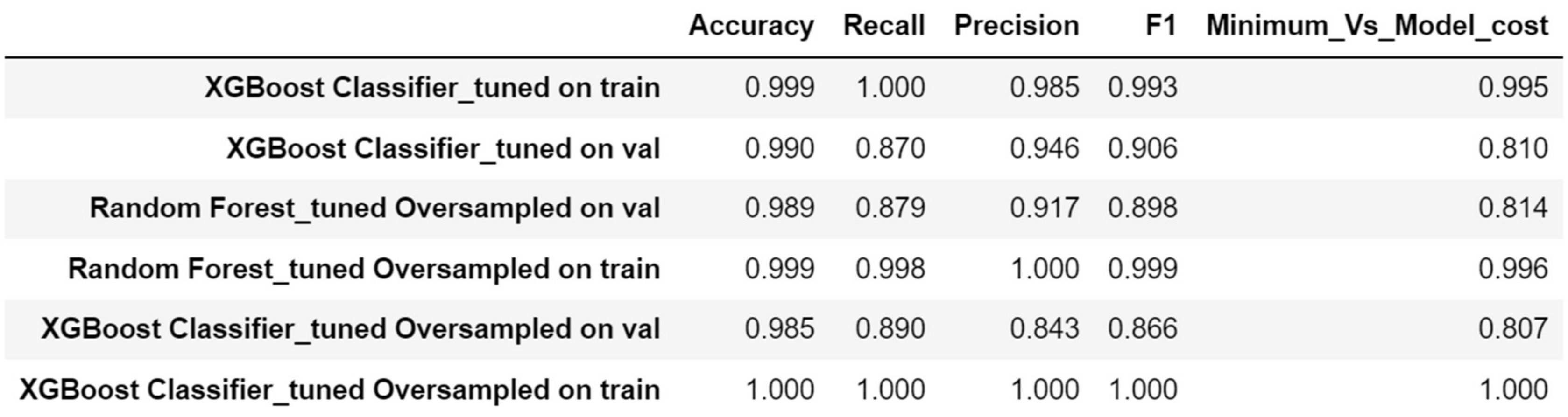
Model Performance Summary

1. I created 6 classification models with original Train data: logistic regression, decision trees, random forest, bagging classifier and boosting methods.
2. And then built 6 classification models using oversampled Train data.
3. Also did the same, using undersampled Train data.
4. Finally chose 3 best performing models among all the models built previously to further tune them to improve the performance.

XGBoost Classifier, Random Forest Oversampled data and XGBoost Classifier Oversampled data are my chosen models because all scores are good and acceptable in train and validation data especially recall and Minimum_Vs_Model_cost.

Model Performance Summary

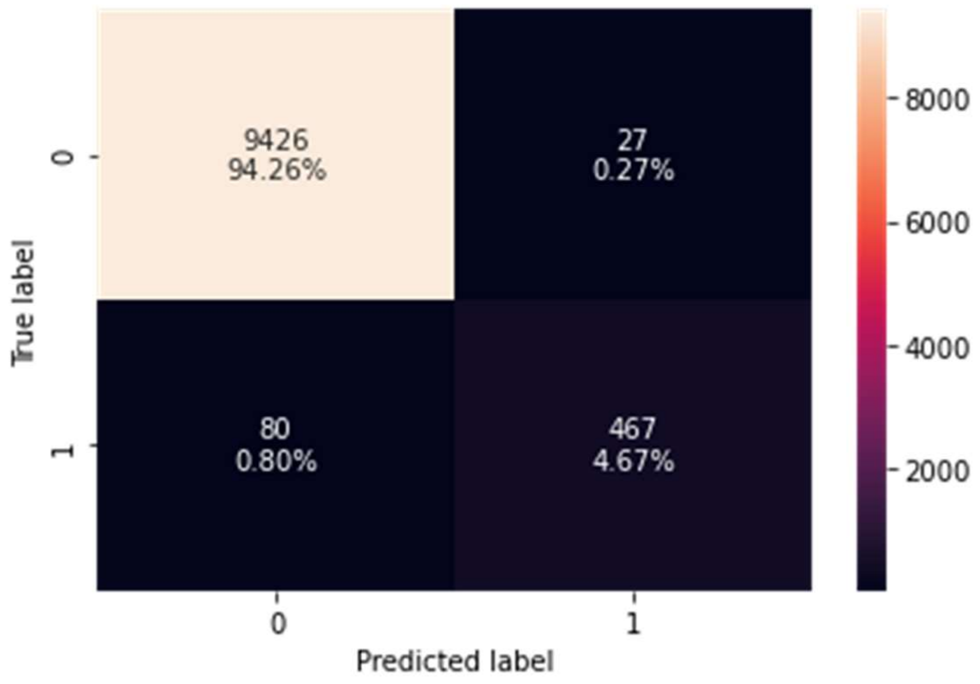
- The below table is about models' performance scores after tuning them.
- between all three tuned model I choose XGBoost Classifier_tuned. The scores are all good, but I think `XGBoost Classifier_tuned Oversampled` has overfitting. In compared to `Random Forest Tuned Oversampled ` all scores in `XGBoost Classifier Tuned` on validation and train are better totally.



	Accuracy	Recall	Precision	F1	Minimum_Vs_Model_cost
XGBoost Classifier_tuned on train	0.999	1.000	0.985	0.993	0.995
XGBoost Classifier_tuned on val	0.990	0.870	0.946	0.906	0.810
Random Forest_tuned Oversampled on val	0.989	0.879	0.917	0.898	0.814
Random Forest_tuned Oversampled on train	0.999	0.998	1.000	0.999	0.996
XGBoost Classifier_tuned Oversampled on val	0.985	0.890	0.843	0.866	0.807
XGBoost Classifier_tuned Oversampled on train	1.000	1.000	1.000	1.000	1.000

Model Performance Summary

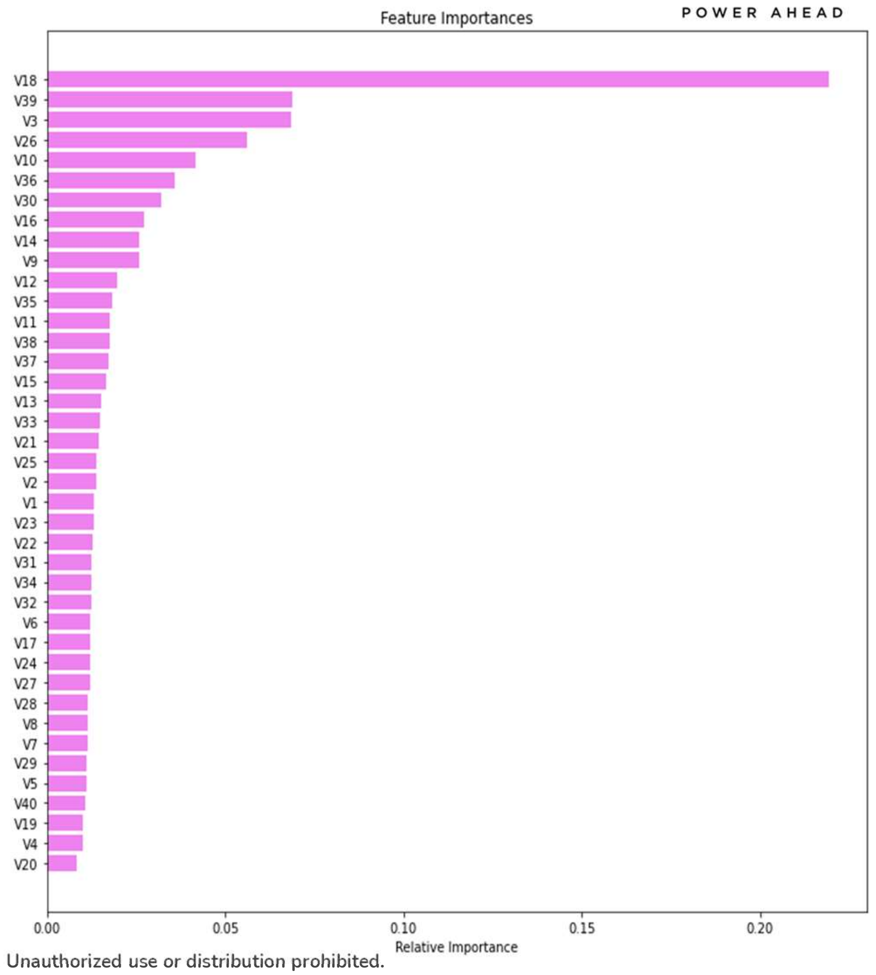
- I set the best final model on the Test data; the confusion matrix results is:
- True positives (TP) that are failures correctly predicted by the model is 4.67%
- False negatives (FN) that are real failures in the generator of wind turbine where there is no detection by the model is 0.8%
- False positives (FP) that are failure detections in the generator of the wind turbine where there is no failure is 0.27%
- True negative (TN) that are real no failures correctly predicted by the model is 94.26%



Model Performance Summary

- V18 is the most important variable in predicting maintenance followed by V39, V3, V26 and V10.

	Accuracy	Recall	Precision	F1	Minimum_Vs_Model_cost
0	0.989	0.854	0.945	0.897	0.794



Business Insights and Conclusions

- We have been able to build a predictive model:

a) that the ReneWind company can deploy this model to identify generator's failures so that the generator could be repaired before failing/breaking and the overall maintenance cost of the generators can be brought down.

b) that the ReneWing company can use to find the key causes that drive maintenance.

c) based on the model, companies can take appropriate actions to build better retention policies maintenance costs.

- Factors that drive the maintenance = V18, V39,V3,V26 and V10
- The model can maximize the ratio of minimum possible maintenance cost and the maintenance cost associated with the model (Minimum_Vs_Model_cost) = 0.79

greatlearning
Power Ahead

Happy Learning !