# WiDS 2021

Feature Engineering to Improve Performance

Natalie Pirkola
MI, USA
npirkola@gmail.com

Stacy Forsyth
IBM, MI, USA
stacyjolly@gmail.com

Elena Barbulescu
MI, USA
e2barbulescu@gmail.com

Kate Tereshchenko
IBM London, UK
kateryna.tereshchenko@gmail.com

## ABSTRACT

The 2021 Women in Data Science WiDS Datathon applied machine learning to the modified version of the Global Open-Source Severity of Illness Score (GOSSIS) Consortium data. This international data set includes the data collected from the first 24 hours during an intensive care unit (ICU) stay and was used to predict whether or not the patient has diabetes mellitus (diabetes). Using a tree-based model, LightGBM, our team focused on data understanding to derive meaningful new features to increase model accuracy.

## KEYWORDS

Diabetes, LightGBM, Machine Learning, WiDS

## 1 Background

Type 2 diabetes mellitus (diabetes) accounts for 90–95% of all cases of diabetes.[1] Individuals have an insufficient insulin production (relative deficiency) and increased resistance to insulin.[1] Insulin is secreted by the pancreas into the blood. Insulin allows sugar to move from blood to be stored or used by cells in the body. Patients with type 2 diabetes are slower to remove sugar from the blood compared to individuals without diabetes resulting in higher blood sugar.[1] Key risk factors for diabetes include older age and elevated BMI.[1] Age is a major risk factor for diabetes and all adults over age 45 should be screened.[1] In general, BMI ≥25 kg/m$^2$ is a risk factor for diabetes, however, the cut point is lower for Asian American populations.[1] Higher risk for diabetes is associated with HIV, history of gestational diabetes, hypertension or dyslipidemia, and polycystic ovary syndrome.[1] Risk is also higher for certain racial/ethnic subgroups including African American, American Indian, Hispanic/Latino, and Asian American.[1] Importantly, approximately one-quarter of people with diabetes and nearly half of Asian and Hispanic Americans with diabetes are undiagnosed.[1] Diabetes is diagnosed when one of the four scenarios is present: blood sugar ≥126 after 8 hours without eating, blood sugar ≥200 after a 2-hour oral glucose tolerance test, an A1C ≥6.5%, or random blood sugar ≥200 mg/dL with symptoms.[1] See Figure 1.

| |
|---|
| FPG ≥126 mg/dL (7.0 mmol/L). Fasting is defined as no caloric intake for at least 8 h.* **OR** |
| 2-h PG ≥200 mg/dL (11.1 mmol/L) during OGTT. The test should be performed as described by WHO, using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water.* **OR** |
| A1C ≥6.5% (48 mmol/mol). The test should be performed in a laboratory using a method that is NGSP certified and standardized to the DCCT assay.* **OR** |
| In a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, a random plasma glucose ≥200 mg/dL (11.1 mmol/L). |
| DCCT, Diabetes Control and Complications Trial; FPG, fasting plasma glucose; OGTT, oral glucose tolerance test; WHO, World Health Organization; 2-h PG, 2-h plasma glucose.<br>* In the absence of unequivocal hyperglycemia, diagnosis requires two abnormal test results from the same sample or in two separate test samples. |

**Figure 1: Diagnosis of Diabetes. Adapted from American Diabetes Association[1]**

Machine learning models have been successfully used to make predictions about patients in ICUs. Many studies have used machine learning to predict the risk of complications in ICU patients. In a review of 258 published studies using machine learning to predict mortality, prognosis, or subpopulations classification, the mean performance measure (AUC) was 0.83 for small studies including fewer than 10,000 patients and 0.94 for large studies with more than 100,000 patients.[2] The two most common machine learning algorithms since 2015 have been support vector machine and random forests, a tree-based model. Other Kaggle competition winners utilized tree-based models. We tried various models and had best results with LightBGM. Therefore, we chose to use LightGBM, a tree-based model, for our final submission.

Decision tree algorithms like LightGBM can be used to solve classification or regression problems.[3] Tree-based models are considered robust to outliers and require less manipulation of data before the model can be applied to the dataset.[4] Tree-based models scale well to large datasets and are able to address complex decision boundaries due to their hierarchical structure.[4] Unlike other decision tree algorithms which split the tree by level or depth, LightGBM splits the tree leaf wise with the best fit.[5] See Figure 2. LightGBM selects the leaf with the max delta loss to grow.[6] This allows LightGBM to reduce more loss than the level-wise algorithm and often provide better accuracy.[5] LightGBM also handles categorical features well compared to other decision tree models.[7] The data supplied included a large number of categorical features, underscoring the value of LightGBM.

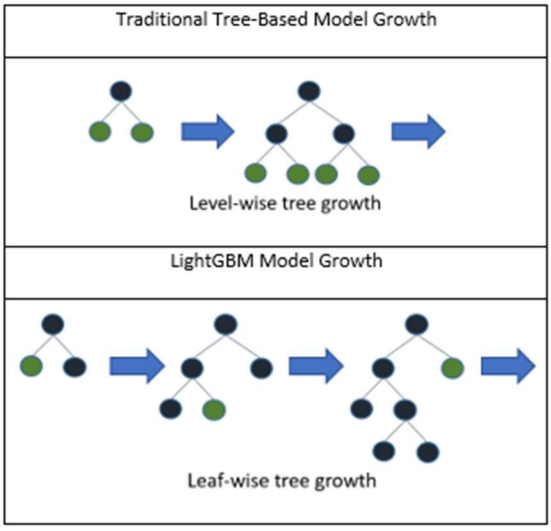


**Figure 2: Model Growth Comparison. Adapted from Microsoft[6]**

## 2 Data Exploration and Preparation

The data set[8] is a modified Global Open-Source Severity of Illness Score (GOSSIS) Consortium data set.[9] The GOSSIS data set was created to develop risk scoring systems for critically ill patients internationally.[9] It includes critical care data from intensive care units in Argentina, Australia, New Zealand, Bangladesh, India, Nepal, Sri Lanka, Brazil, and the United States.[9] The data set supplied included 181 feature columns and 130,157 patient encounters where each row is a unique encounter, reflecting a 24-hour period of an individual in the ICU.[9]

Exploratory data analysis was conducted using SweetViz, seaborn, pandas, matplotlib and a review of other publicly available participant notebooks (see Acknowledgements). Data was assessed for completeness. Missing data was predominantly filled with the mean of the feature values, although each feature was individually assessed for the most appropriate fill option for the statistical spread of the values as well as the resultant impact on the final model performance. Categorical features were encoded. Continuous features were scaled and normalized, as necessary.

The data set had an imbalanced class distribution with the minority class of a patient having diabetes being just over 20%. After experimenting with resampling the dataset and generating synthetic samples via SMOTE, a combination of under sampling and over sampling approach was adopted.

Exploratory analysis was conducted on binning various continuous values. Most commonly, the continuous data was retained for analysis. Notable exceptions include the binning of various blood pressure values to indicate stages of hypertension control for people with diabetes. Reported body mass index (BMI) was calculated from height and weight and compared to reported BMI if available. Height and weight values had fewer missing values in the data than BMI values. Also, it was considered more reliable by the subject matter expert to impute the values for height and weight rather than BMI itself. Therefore, calculated BMI was used for the analysis.

## 3 Feature Engineering

Feature engineering included the use of Octopus ML as well as subject matter expertise to create new features to include in the model. Additional features were created by calculating the results of other features and iterating. Trial and error through iterations allowed us to determine strategies for feature engineering and model tuning. Although not all iterations improved the score, we learned from each iteration. Overall, more iterations contributed to our learning about adjustments to variables, resulting in a higher the score over time. The Glasgow Coma Scale (GCS) components were available in the data set (‘gcs_motor_apache', 'gcs_verbal_apache', 'gcs_eyes_apache'). The GCS combined score was calculated and the results were binned according to the GCS methodology for scoring. BMI was then concatenated with gender and ethnicity to increase feature significance. Model performance improved by using this merged feature over the individual features alone which we found through iterating combinations of these variables

## 4 Results

The LightGBM model achieved a ROC AUC near 0.87 indicating a strong performance predicting diabetes in the ICU patient with 24 hours of data, however, performance could be further enhanced. ROC AUC performance was a mix of known test data and blinded test data. As tree-based models tend to overfit, tuning the model to reduce overfitting and applying to another dataset could improve performance. Adjusting the max depth parameter in the LightGBM model can reduce overfitting.[6] Similarly, the number of leaves could be adjusted to reduce the complexity of the model, thereby reducing overfitting (Bahmani, 2021).[7]

The model used the default gradient boosted decision trees version of LightGBM which has more memory issues.[7] DART gradient boosting could be used if over-specialization was an issue.[7] Performance could also be assessed using Gradient-based One-Side Sampling as the boost method.[7] This data set was also unbalanced. Adjustments were made to sampling; however, LightGBM also has a parameter to handle unbalanced data. Adjusting the model to run as an unbalanced data set may also impact performance.[7] LightGBM can also be adjusted to scale for the weight of the number of positive examples (cases labelled with diabetes) compared to negative examples (cases without diabetes). Comparing methods to adjust for the unbalanced data set could also improve performance.[5]

## 5 Conclusion

The promising ROC AUC score of 0.87 from our modeling indicates that machine learning is a viable method to identify patients with diabetes early in their stay in the ICU. Identifying patients with diabetes early in the ICU stay can potentially improve their clinical care provided by better understanding the treatment a patient may need for the diabetes as well as improve the treatment of other common comorbidities associated with diabetes. Hospitals can consider using an algorithm like this to identify diabetes quickly and accurately to improve ICU staffing, patient risk prediction, and coding processes for payment. The model was successful in predicting patients with diabetes, using only data available in the first 24 hours of an ICU stay. Providing additional data from the hospital or ICU stay could strengthen the predictive value. Providing additional data elements, such as other comorbidities and lab values, would further enhance the predictive capabilities.

This LightGBM model was based on a modified version of the Global Open-Source Severity of Illness Score (GOSSIS) Consortium data set[9] which indicates that it could be applicable internationally as well as in the United States. As assessed by ROC AUC, the ability to accurately predict patients with diabetes was strong. However, predictions should be combined with clinical judgment before affecting patient treatment decisions. A false positive result, predicting diabetes when it does not exist, may cause a hospital to assign more staff to monitoring a patient blood sugar without cause. A false negative, predicting no diabetes when it does exist, may results in under-monitoring and undertreatment of an individual. Overall, the goal of the competition was to provide a model to identify patients with diabetes without patient or family input. If the model predicts diabetes, it could be used to prompt additional questions from the patient, family, or external data sources for verification. Furthermore, prior to implementing this model in a clinical tool or for clinical decision making, this analysis should be replicated with another robust data set. Our results, as well as results from other WiDS competition models, show promising implications for the application of machine learning to enhance diabetes identification in the ICU to improve patient treatment and outcomes.

## ACKNOWLEDGMENTS

An end-to-end Data Science Methodology framework session and mentoring was delivered by Erika Agostinelli (IBM London, UK)

Appreciation goes to Kaggle contributor Siavash. (https://www.kaggle.com/siavrez) for sharing high scoring code during the competition which allowed us to focus on additional feature engineering. Thank you to previous WiDS competition winners who supported Siavash's helpful code including: Kain's work (5 place), Dan Ofer's works (1 place), jayjay75's work (3 place) and dynamic24's write-up (6 place).

## REFERENCES

1. American Diabetes Association. (2021, January). Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care*, *44* (Supplement 1), S15-S33. Retrieved from https://doi.org/10.2337/dc21-S002
2. Shillan, D., Sterne, J., Champneys, A., & Gibbison, B. (2019, August 22). Use of Machine Learning to Analyse Routinely Collected Intensive Care Unit Data: A Systematic Review. *Critical Care, 23*(1), 284.
3. Varghese, D. (2018, December 6). *Comparative Study on Classic Machine learning Algorithms.* Retrieved from Towards Data Science: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222
4. Elite Data Science. (2019). *Modern Machine Learning Algorithms: Strengths and Weaknesses.* Retrieved from Elite Data Science: https://elitedatascience.com/machine-learning-algorithms
5. Kasturi, S. (2019, July 11). *XGBOOST vs LightGBM: Which algorithm wins the race !!!* Retrieved from Towards Data Science: https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d
6. Microsoft. (2021). *Features.* Retrieved from LightGBM: https://lightgbm.readthedocs.io/en/latest/Features.html
7. Bahmani, M. (2021, March 12). *Understanding LightGBM Parameters (and How to Tune Them)*. Retrieved from Neptune Blog: https://neptune.ai/blog/lightgbm-parameters-guide
8. Stanford University WiDS Worldwide team. (2021, January 5). *WiDS Datathon 2021: Data*. Retrieved from Kaggle : https://www.kaggle.com/c/widsdatathon2021/data
9. MIT Critical Data Laboratory . (2021). *GOSSIS: About Us.* Retrieved from GOSSIS: https://gossis.mit.edu/about/