Team Final Project
Data Science for Business
DECISION 520Q
Professor Alex Belloni
Oct 13, 2024

MQM 25 Summer Sec C Team 55
Lance Chi (bc335)
Sandro Chumashvilli (sc967)
Vidhi Jain (vj81)
Prerana Munipalli (pm300)
Ruoyi Xiao (rx61)

Win-Win Healthcare: Enhancing Stroke Prevention and Novo Nordisk's Impact Through Glucose Precision

Business Understanding:

Novo Nordisk is one of the biggest insulin producers globally and has made its mark through innovative research, clinical trials, and strong partnerships with healthcare providers. With recent insulin price caps shaking up the market, it's more important than ever for the company to use data-driven insights to help doctors adopt their products. This project, which looks at the link between glucose levels and stroke risk, gives a chance to go beyond just pricing and bring extra value to doctors, patients, and healthcare systems, reinforcing its position as a leader in diabetes care.

Leveraging Data Science to Enhance Novo Nordisk's Market Position

Novo Nordisk has built a strong influence with healthcare providers through its research-driven approach to diabetes treatment, developing therapies like insulin analogs (NovoRapid, Levemir) and GLP-1 receptor agonists (Victoza). This research has earned doctors' trust in prescribing its safe and effective products. However, U.S. insulin price caps and competition from biosimilars now require Novo Nordisk to emphasize the added value of its treatments, particularly in preventing complications like strokes. Key trials like DEVOTE and LEADER highlight the benefits of Novo Nordisk's drugs, while the company's move toward precision medicine and machine learning strengthens its market position.

Machine Learning and Its Role in Driving Novo Nordisk's Strategy

The **Emerging Risk Factors Collaboration (ERFC)** study, published in *The Lancet* in 2010, analyzed data from 102 prospective studies involving over 698,000 individuals to investigate the relationship between

blood glucose levels and cardiovascular diseases, including stroke. The study found that individuals with diabetes had a 2.27-fold increased risk of ischemic stroke compared to non-diabetics, indicating a strong link between high glucose levels and stroke risk. This large-scale analysis highlighted how elevated glucose levels contribute to vascular damage, leading to an increased likelihood of stroke. Our project wants to build upon it and employ machine learning to analyze patient data, particularly glucose levels, to predict stroke risk. By identifying patients at high risk due to poor glucose control, we could provide doctors with data-driven insights that inform their prescription choices, demonstrating the efficacy of Novo Nordisk's drugs and aiding in the prevention of severe complications.

With stroke prevention being crucial for diabetes patients, our research could position the company's insulin and glucose-regulating products as essential. This is important as new competitors enter the market, and the company must enhance its offerings beyond just products. By delivering targeted data to healthcare providers, Novo Nordisk can boost *drug adoption rates* and support informed treatment decisions, maintaining a competitive edge in a market where pricing is no longer the main differentiator.

Strategic Partnerships with Doctors and Healthcare Systems

Through sales representatives and digital health platforms, the company maintains ongoing communication with healthcare providers, sharing the latest research, clinical trial results, and patient care tools. As the market evolves due to the insulin price cap, these relationships will grow even more valuable. By highlighting the benefits of prescribing Novo Nordisk's drugs—such as reducing stroke risk in high-risk patients—the company ensures its products remain central to doctors' treatment plans.

The Role of Price Caps and the Future of the Insulin Market

The recent U.S. insulin price cap under the Inflation Reduction Act limits Medicare patients to $35 per month, forcing companies like Novo Nordisk to cut prices by up to 75%, impacting profit margins. To stay competitive, the company must shift focus from pricing to value-added services and leverage its research. Using machine learning to identify high-risk patients for complications like strokes can help doctors prescribe drugs more effectively, strengthening the company's role in patient-centered care.

Data Understanding:

The dataset, widely recognized for credibility on Kaggle, was created by combining real-life patient profiles from different regions. While Fedesoriano, the author of the database, also created several other datasets on illness (Heart failure, Hepatitis C, etc.) by combining real-life patient profiles from different regions and countries (for example, Cleveland (state) and Switzerland(country)), the source for our dataset remains confidential. More information on this data is not available and is needed for confidence in this project. The dataset consists of 5,110 records with 12 variables, including BMI, glucose levels, smoking status, work type, and hypertension. The target variable is stroke (0 = no stroke, 1 = stroke), making this a supervised learning problem. A significant class imbalance exists, with only 249 stroke cases, which could lead to high false negatives. Key variables include glucose levels, ranging from 55 to 272, potentially introducing scaling bias, and BMI, a crucial stroke risk factor, where unstandardized values for ages 0.08 to 82 may cause bias. Additionally, 30% of the records have “Unknown” smoking status, which could introduce selection or imputation bias depending on how it is handled. We addressed these issues during data training to mitigate bias risks.

Data Preparation:

For data cleaning, we first removed irrelevant ID columns and filtered out rows with "Other" gender labels. Categorical variables, like gender, were converted into binary form (males as 1, females as 0). Using the fastDummies package, we transformed other categorical variables, such as work_type, residence_type, and smoking_status, into dummy variables, creating separate binary columns for each category. For BMI, missing values were imputed based on the mean BMI within the same average glucose level group. Continuous variables, such as BMI and avg_glucose_level, were categorized into ranges taken from the World Health Organization for easier analysis. To handle missing smoking status values, we applied a multinomial logistic regression model for prediction, improving accuracy. The age variable was standardized into a z-score, ensuring all numeric variables were on the same scale, and individuals under 18 years of age were removed, as most strokes in minors are due to congenital abnormalities.

Modeling

We seek to develop a predictive model that offers actionable insights for medical and pharmaceutical stakeholders so Novo Nordisk, Doctors and most importantly - patients can benefit from our project.

Random Forest:

This is an ensemble learning method based on decision trees. We trained a Random Forest classifier to predict stroke occurrence by generating multiple decision trees and aggregating their predictions. The Random Forest model allows us to evaluate the importance of various features, such as age, glucose levels, and other health factors. We also applied hyperparameter tuning to optimize the model using techniques like cross-validation to determine the best number of trees (ntree) and the best number of features to split at each node (mtry).

XGBoost:

XGBoost builds a series of decision trees where each new tree corrects the errors made by the previous ones. We used this model to predict stroke occurrences as well. Similar to the Random Forest model, we optimized the parameters through hyperparameter tuning, adjusting settings like the learning rate, tree depth, and subsampling ratio to find the best-performing model. XGBoost's strength lies in its handling of imbalanced data, regularization to prevent overfitting, and efficiency in computation.

Logistic Regression:

We used logistic regression to model the relationship between the dependent variable (stroke) and several independent variables (such as glucose levels and age). Logistic regression provides us with interpretable coefficients in terms of odds ratios, allowing us to assess the likelihood of a stroke based on specific risk factors.

1. Imbalanced Dataset and the Class Imbalance Problem

As mentioned beforehand, the dataset initially showed a significant imbalance, with 4,007 cases of individuals without a stroke compared to only 247 stroke cases. This poses a challenge, as most machine learning algorithms tend to favor the majority class, risking frequent predictions of "no stroke" and overlooking critical stroke cases. To address this, we applied SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic examples of the minority class—in this case, stroke cases—by

interpolating between existing data points. This helped balance the dataset, allowing the model to be exposed to more stroke cases and learn from them effectively. By balancing the data, we ensured the model had a fair chance to detect stroke cases..

2. Data Splitting and Cross-Validation

We divided the dataset into 80% for training and 20% for testing. The training set enables the model to learn patterns, while the test set assesses its ability to generalize to new data. Additionally, we implemented 5-fold cross-validation during training, where the training data is split into five parts, with the model trained on four parts and tested on the remaining one. This technique helps prevent overfitting, ensuring the model performs well on unseen data.

We employed **XGBoost**, a high-performance gradient boosting algorithm, to predict stroke risk. Think of it as a group of experts working together to make the best prediction. Each expert (or decision tree) tries to fix the mistakes of the one before, so the final result is much more accurate. XGBoost constructs decision trees sequentially, each one learning from the errors of the previous, resulting in better prediction power than Random Forest in many cases. Our model trained on 23 features, such as glucose levels, age, and hypertension, achieved strong results with a balanced accuracy of 74.9%, slightly outperforming Random Forest's 73.8%. XGBoost's specificity, in identifying stroke cases, was 52.04%, compared to Random Forest's 50%.

Model Performance Overview:

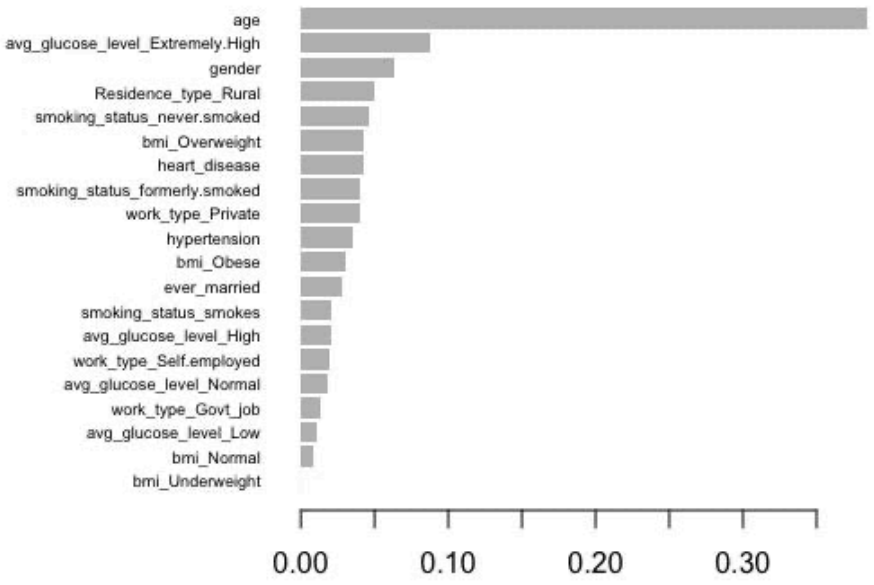
- **Accuracy:** XGBoost outperformed Random Forest in stroke detection, with better specificity for high-risk individuals.
- **Sensitivity & Specificity:** XGBoost more effectively balanced identifying stroke vs. non-stroke cases, particularly improving specificity, or the ability to detect actual stroke cases.
- **AUC (Area Under Curve):** Although the AUC values for XGBoost and Random Forest are close and may vary slightly in different runs due to the stochastic nature of the models, XGBoost consistently provides competitive discriminatory power.

Feature Importance:

Similar to Random Forest, XGBoost ranks features by importance, with glucose levels identified as a key

predictor of stroke risk. However, correlation does not imply causation; factors like age and hypertension may drive this relationship. For example, ice cream sales may rise alongside drowning incidents in summer due to heat rather than a causal link.

These refinements would enable us to provide even more accurate predictions and deeper insights into the relationship between glucose levels and stroke risk. Based on the model's results, one of the key findings is



the significant role of **extremely high glucose levels** in predicting stroke risk. As seen in the feature importance chart, **extreme glucose levels** are the second most critical factor after age. This underscores the strong association between abnormal glucose levels and stroke occurrence.

To put it simply, consider glucose as the fuel for your body. When levels are too high, it can overload the system, much like an engine overheating from excess fuel. The model shows that individuals with very high glucose levels are at significantly greater risk for stroke. This is understandable, as persistently high glucose can damage blood vessels over time, raising the chances of blockages or ruptures—major factors that lead to strokes.

This finding is significant because it has practical implications for healthcare providers and companies like Novo Nordisk. It emphasizes that **controlling glucose levels**, especially in individuals with dangerously high levels, is crucial. Medications that regulate glucose can help lower the risk of severe events, such as strokes. While the model doesn't establish causation, it underscores the importance of glucose management as a targeted intervention for at-risk populations. Thus, controlling glucose levels may be a key strategy in reducing stroke risk across patient groups.

Evaluating the models - this will help us understand how well machine learning models predict outcomes, like stroke risk, and XGBoost consistently proves to be the better choice. The AUC-ROC (Area Under the

Curve - Receiver Operating Characteristic) measures how well a model distinguishes between stroke and non-stroke cases. XGBoost, with a score of 0.8953, outperforms Random Forest's 0.8879, meaning XGBoost does a better job of sorting patients into the correct categories—similar to a more accurate sorting system.

The Precision-Recall AUC evaluates the model's ability to identify true stroke cases, which is crucial when strokes are rare. While both models have similar scores here, XGBoost excels due to its higher recall, making it more reliable for identifying actual stroke cases. Think of it like a more sensitive metal detector—XGBoost is better at catching potential risks.

Precision measures how accurate the model is when it predicts a stroke. Random Forest has perfect precision (1.0), but it only identifies 18% of real stroke cases (low recall). On the other hand, XGBoost's recall is 45%, meaning it catches more real stroke cases. In healthcare, this is critical: XGBoost acts like a security system that identifies more potential risks, preventing more dangerous situations.

Finally, the F1-Score balances precision and recall, with XGBoost scoring significantly higher (0.6154) than Random Forest (0.3103). This makes XGBoost the better model for medical applications where catching as many stroke cases as possible is essential for early intervention and prevention.

Deployment

Novo Nordisk can take advantage of its existing partnerships with healthcare providers to roll out this predictive model, much like IBM Watson was used by Memorial Sloan Kettering to assist doctors in identifying optimal cancer treatments. This approach could involve embedding the stroke risk model into digital health platforms already used by doctors for patient management. For instance, an integration within **Epic Systems**, the most widely used electronic health record (EHR) platform in the U.S., could enable real-time stroke risk predictions based on patient glucose levels.

Doctors would receive alerts about high-risk patients during routine check-ups, allowing them to adjust treatment protocols immediately. For example, a physician reviewing a diabetic patient's data could see a flagged alert stating, "This patient's glucose levels indicate a 35% higher stroke risk based on predictive modeling." The doctor could then recommend immediate interventions, such as tighter glucose control using

Novo Nordisk products, or lifestyle changes. This mirrors how Google Health partnered with hospitals to use AI for detecting breast cancer, integrating seamlessly into clinical workflows.

Pharmaceutical Deployment: Novo Nordisk could also embed these insights into their sales representatives' tools. Representatives can use the predictive model to customize conversations with healthcare providers, focusing on data-driven prevention strategies. Imagine a representative discussing with a physician: "Doctor, our predictive tool has identified a significant correlation between high glucose levels and stroke risk in your patient population. Here's how Novo Nordisk's treatments can help mitigate that risk." This shift from purely product sales to data-driven insights would deepen relationships with doctors and position Novo Nordisk as a trusted partner in patient care.

Patient Engagement: The model could also be deployed through patient-facing mobile apps or wearables, similar to how Apple's Health app integrates heart rate monitoring with predictive insights for atrial fibrillation. Novo Nordisk could create a patient app that alerts individuals with diabetes when their glucose levels reach critical thresholds, warning them about their stroke risk. For instance, "Your glucose levels have exceeded 200 mg/dL—your stroke risk has increased by 20%. Please contact your healthcare provider for guidance." This personalized engagement empowers patients to take proactive steps in their care.

Ethical Considerations: Ensuring transparency is essential. Patients and doctors should understand that while the model can flag risks, it is not a replacement for medical judgment. Ethical concerns around patient autonomy and data privacy must be addressed, especially under laws like HIPAA. Novo Nordisk should ensure that all predictions are interpretable and explainable, similar to the principles employed in explainable AI models used by companies like FICO in financial services.

Risks and Challenges

Data Risks:

The lack of transparency in the dataset's source raises concerns about its reliability, much like Google Flu Trends' failure due to poor data. Missing values (e.g., 30% "Unknown" smoking status) may skew results, while categorizing continuous variables like glucose and BMI risks losing valuable details. These issues

could lead to unreliable predictions in medical settings, making the model less useful for general patient populations.

Modeling Risks:

Overfitting remains a concern, as XGBoost may perform well on training data but poorly on real-world data. While SMOTE helped balance stroke cases, it risks introducing artificial patterns that don't exist in real patients, leading to inaccurate predictions. Additionally, with XGBoost's recall at 45%, more than half of stroke cases could be missed, potentially undermining patient safety.

Deployment Risks:

Over-reliance on the model poses ethical risks, especially given its imperfect recall—missing stroke cases could be fatal. One critical issue is data interoperability. Healthcare data is notoriously fragmented across various platforms, and integrating predictive models into EHRs or mobile apps could face resistance from providers due to technical complexity or perceived workflow disruptions. To overcome this, Novo Nordisk could partner with **HL7 International**, the organization responsible for healthcare data standards, to ensure smooth integration of their model into various systems without compromising workflow efficiency.

Mitigating Assumptions:

High precision in the model might lead to missed diagnoses of borderline cases, which could have severe consequences in healthcare. While XGBoost outperformed Random Forest, simpler, more interpretable models like logistic regression may be better suited for gaining doctors' trust, as complex models may hinder clinical acceptance.

In summary, data quality, overfitting risks, and deployment challenges must be addressed to ensure this model's success in real-world healthcare applications. Careful evaluation, transparent communication, and integration into existing workflows are essential to mitigate these risks.