DBA5106 Group Project

Predictive Modeling to Identify Diabetic Patients

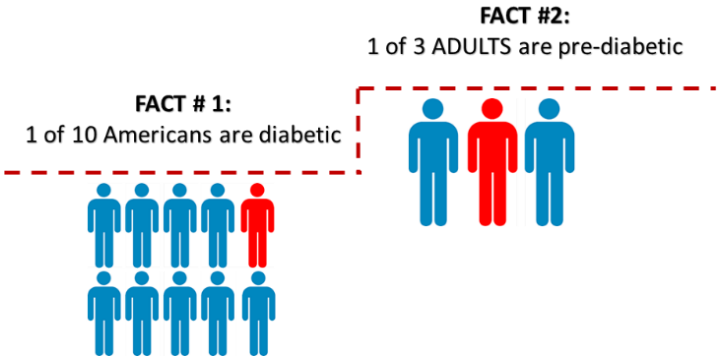
AUTHORS (Team 16):

Gino Martelli Tiu (A0231956Y)
Susan Koruthu (A0231905L)
Widya Salim (A0231857Y)
Xhoni Shollaj (A0231930N)

Section 1: Introduction

Section 1.1: Diabetes has significant economic and productivity costs.

As of 2019, the World Health Organization (WHO) ranks diabetes as one of the top 10 leading causes of death with over 1.5 million deaths per annum globally. In the United States alone, the American Diabetes Association estimates a staggering $237 billions in direct medical costs and another $90 billions in lost productivity to diabetes[1]. This roughly means that 1 in every $4 healthcare dollars is spent on diabetes treatment.



The brunt of medical costs is borne by insurance outfits[2], as shown below:

- Government insurance (67%)
- Private insurance (31%)
- Uninsured (2%)

Section 1.2: 95% of diabetes cases are reversible or preventable.

Characterized by sustained high levels of sugar over time, the disease is caused when the body fails to produce enough insulin or is unable to respond to it properly – leading to other medical complications affecting the heart, blood vessels, nerves, kidney, cognitive function and increasing a person's mortality rate when infected with viruses. There are generally 3 types of diabetes[3]:

- Type 1 (Auto-immune diabetes)
- Type 2 (Preventable diabetes)
- Pre-diabetes (Reversible diabetes)

---

[1]American Diabetes Association. (n.d.). Economic Costs of Diabetes in the U.S. in 2017. *Economic Costs of Diabetes in the U.S. in 2017*, *vol. 41*(May 2018, 41(5)).
[2]American Diabetes Association. (n.d.). Economic Costs of Diabetes in the U.S. in 2017. *Economic Costs of Diabetes in the U.S. in 2017*, *vol. 41*(May 2018, 41(5)).
[3] Centers for Disease Control and Prevention. (June 11, 2020). *What is Diabetes?* U.S. Department of Health & Human Services. https://www.cdc.gov/diabetes/basics/diabetes.html

Of these, type 2 and pre-diabetes account for over 90-95% of total diabetes cases. These 2 types develop over the course of many years and can easily be diagnosed by including inexpensive checks to annual physical exams currently in place. More so, even when diagnosed, diabetes could be reversed or delayed through lifestyle changes.

Section 2: Business Motivation and Scope

Section 2.1: Insurance has a vested interest in mitigating preventable diabetes.

As a business model, health insurance companies charge a monthly or annual fee in exchange for covering all or part of a person's future medical expenses. Premiums and deductibles are then charged based on the insured person's risk attributes and then pooled together to either invest in interest bearing instruments or cover claims as and when they arise. Hence, the lesser the number of claims, the higher the profit is for an insurance company.

In terms of numbers, claims for diabetes treatments incur an average medical expenditure of $16,752 per person per year. This is a 2.3 multiplier relative to non-diabetics. The amount could even further increase by an additional $1,900 if complications occur. Even in a co-pay scenario, the amount of claims is significant.

Section 2.2: Insurance companies can improve public health while increasing profits

Currently, insurance companies commonly screen for diabetes by comparing a claimant's Body Mass Index (BMI), Blood Glucose Level after Two Hour Oral Glucose Test (OGTT) and Glycohemoglobin level against a benchmark guideline provided by the American Diabetes Association. This method however, only identifies around 30% of diabetics.

Our paper argues that insurance companies can move from reactively adjusting premiums to a more active role. With the wealth of information collected during the underwriting and claims process, insurance companies can build a classification model that can:

(1) Compute the probability of positive diagnosis given lifestyle & medical test variables
(2) Identify high risk patients (potential pre-diabetes) for monitoring and intervention
(3) Show insured clients' current state of health relative to medical benchmarks
(4) Directly identified patients to merchant partners who could help them effect required lifestyle changes to prevent, delay or alleviate the after-effects of diabetes.

With these predictive and decision support systems in place, insurers can unlock natural synergies, pad their bottom line while promoting the overall wellbeing of its clients. As an initial proof of concept, the team aims to build models to address points (1) and (2) and tie it into an actionable business proposition for the industry.

The team's effectiveness metric of choice is the expected value (EV) of diabetes expenses claims, computed as follows:

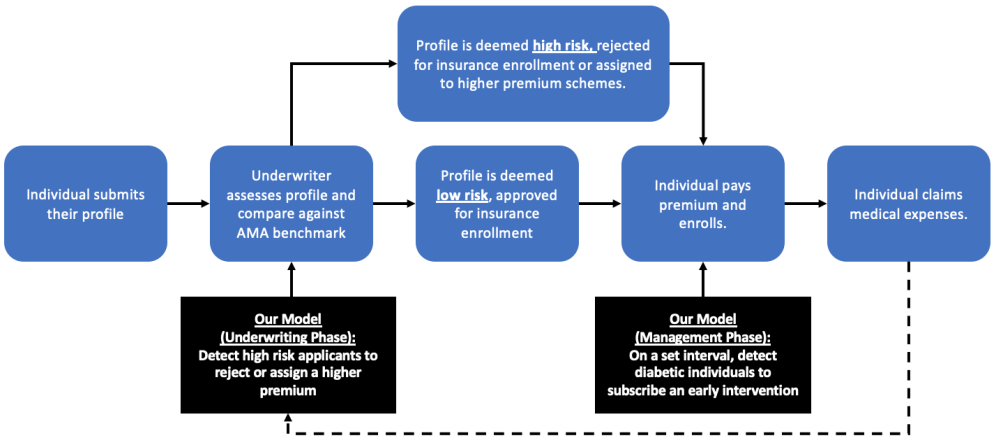
$$Expected\ Value\ of\ Diabetes\ Expenses\ Claims\ =$$
$$P(Diabetic) \times [\ P(Predicted\ Diabetic \mid Diabetic)\ \times \$16,752\ +\ P(Predicted\ NonDiabetic \mid Diabetic) \times \$18,652]$$
$$+\ P(NonDiabetic) \times [\ P(Predicted\ NonDiabetic \mid NonDiabetic) \times \$0\ +\ P(Predicted\ Diabetic \mid NonDiabetic) \times \$200]$$

Insurance companies can use this to determine expected expense claims for a particular group and adjust their premiums accordingly. Compared to the existing detection methodology, the group believes that our model will significantly reduce total expected claims. The below table shows the associated medical cost of diabetes claims.

Probability	Case	Cost	Remark
$P(Predicted\ Diabetic \mid Diabetic)$	True Positive	$16,572	Base cost of diabetes treatment.
$P(Predicted\ NonDiabetic \mid Diabetic)$	False Negative	$18,652	Base cost + cost associated with complications due to the lack of early treatment ($1,900).
$P(Predicted\ NonDiabetic \mid NonDiabetic)$	True Negative	$0	No cost.
$P(Predicted\ Diabetic \mid NonDiabetic)$	False Positive	$200	Cost for diabetes screening test.

As visualized below, the model can be implemented seamlessly during different phases of the insurance cycle. When implemented during onboarding, it can help underwriters eliminate high risk applicants or match them to insurance schemes commensurate with their risk profile.

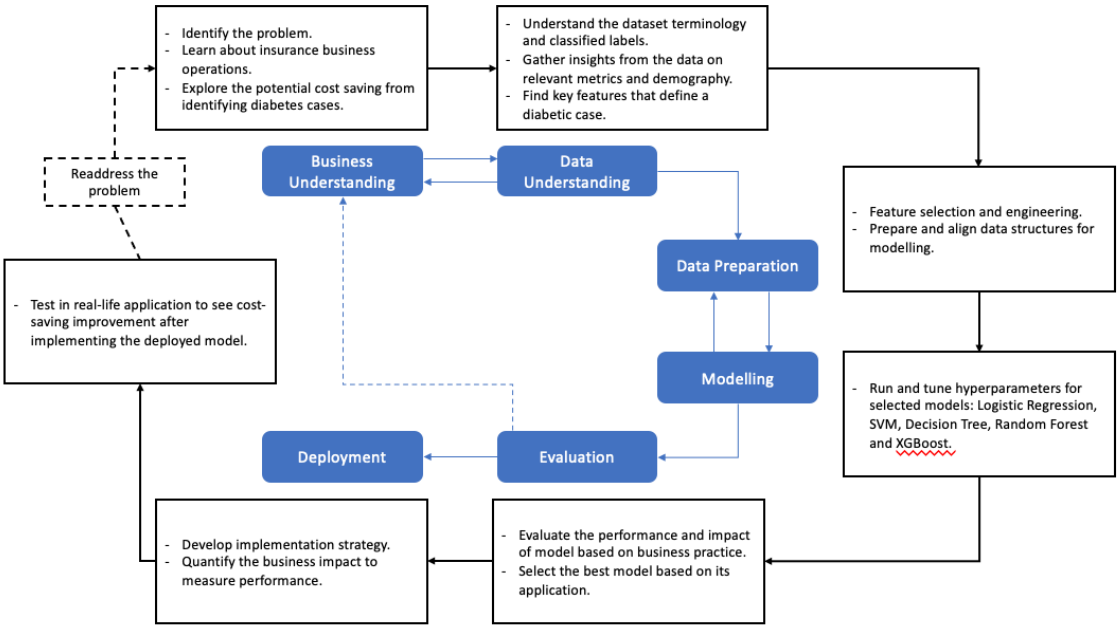
Meanwhile, an implementation for existing clients allows the company to do early detection and intervention thereby reducing future expense claims. This paper will focus on the latter case of early intervention with the goal of reducing claims.



Section 3: Data Handling, EDA and Feature Engineering

Section 3.1: Methodology

The below business analytics workflow shows the end to end thought process followed. Emphasis was placed on understanding the business and potential value add areas before drilling down on the business process where our model would make the most impact.



Section 3.1: Dataset & Feature Engineering

The team referenced the National Health and Nutrition Examination Survey (NHANES) dataset for diabetes, which contains information akin to what insurance firms would have access to throughout the life of the customer relationship.

The original dataset consists of 4 separate tables: (a) demographics, (b) blood mineral, (c) physical examination results and (d) the diabetic/non-diabetic labels. Overall, there were over 83 original features and 10,175 individual rows, which were joined using SEQN (individual id in the dataset) into one table.

Further preprocessing was done to drop duplicate columns, rows with missing values, reducing the observation to 8,971 rows. Columns were likewise relabeled to eliminate abbreviations and complex terminology. The following features were also engineered (*) or regrouped to gain more insights during our EDA:

Feature Name	Type	Description	Remarks

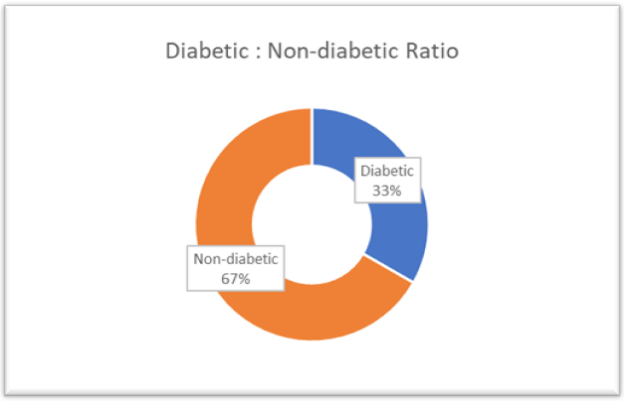
High Risk	Engineered	Individuals with the following characteristics are marked as high risk:
		- BMI ≥ 25
		- Two Hour Glucose(OGTT) (mg/dL) for various age groups:
		  - Age 60: OGTT =110
		  - Age 50: OGTT ≥ 120
		  - Age 40: OGTT= 120
		- Glycohemoglobin (%)' ≥ 5.7
			Identify individuals which are deemed as diabetic/high risk according to existing AMA guidelines

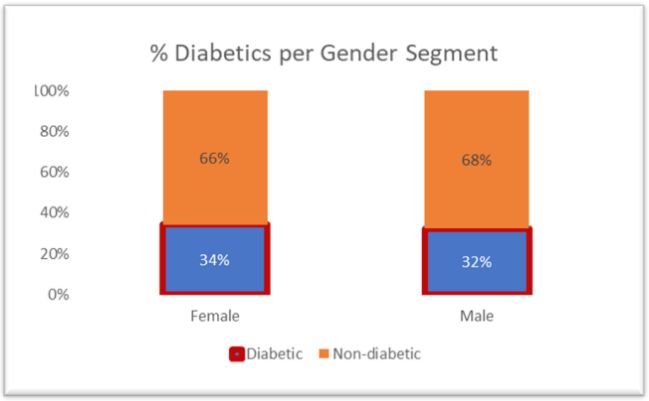
Large Family	Engineered	Indicates whether a household is bigger than 3 people.	Larger families might have different lifestyles.
BMI Group	Engineered	**'Underweight'** when BMI < 18.5 **'Healthy'** when BMI is between 18.5 - 24.9 **'Overweight'** when BMI is between 25 - 29.9 '**Obese'** when BMI > 30	BMI group is one of the main biomarkers of diabetic patients.
Age Group	Engineered	Grouped by decades (e.g: 0 - 10 years, 10 - 20 years, and so on)	Age is one of the leading indicators for diabetes.
College Educated	Engineered	'**High'** when family income > $100K **'Medium'** when family income is between $55K - $100K **'Low'** when family income is < $55K	Higher education could lead to a healthier lifestyle.
Socioeconomic Status	Engineered	'**High'** when family income > $100K **'Medium'** when family income is between $55K - $100K **'Low'** when family income is < $55K	Similar to education, higher socioeconomic status could lead to a healthier lifestyle.
Marital Status	Regrouped	Regrouped to Married, Single, Separated/Divorced or Unknown	Simplify marital status.
US Stay Length	Regrouped	Regrouped to Short Term vs Long Term Stay	Longer stay in the U.S could affect lifestyle habits.
Citizenship	Regrouped	Regrouped to US vs Non US Citizen	U.S citizens might have different lifestyle habits VS. non-U.S citizens due to cultural differences.

Section 3.2: Exploratory Data Analysis (EDA)

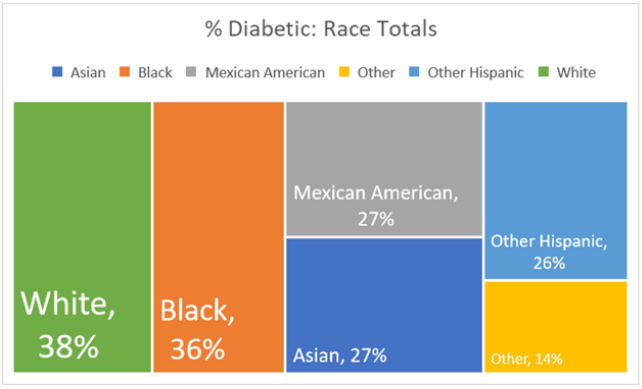
To better understand the relationship between diabetes and other features, our preliminary EDA focused on gaining insights on individuals' demography and medical status which we hypothesised could affect their predisposition to diabetes. A few characteristics which gave us the best insights were gender, race, family income, BMI, annual physical results and age group. Observations and insights for these characteristics are explored in the succeeding subsections.

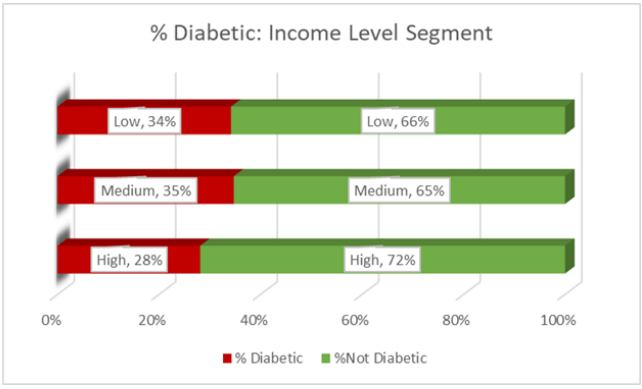
Section 3.2.1: A Demographic EDA View of Diabetes





- Roughly a third of the sampled population is diabetic.
- As a percentage of gender total, females have a slightly larger proportion of positively confirmed diabetes (+2%).

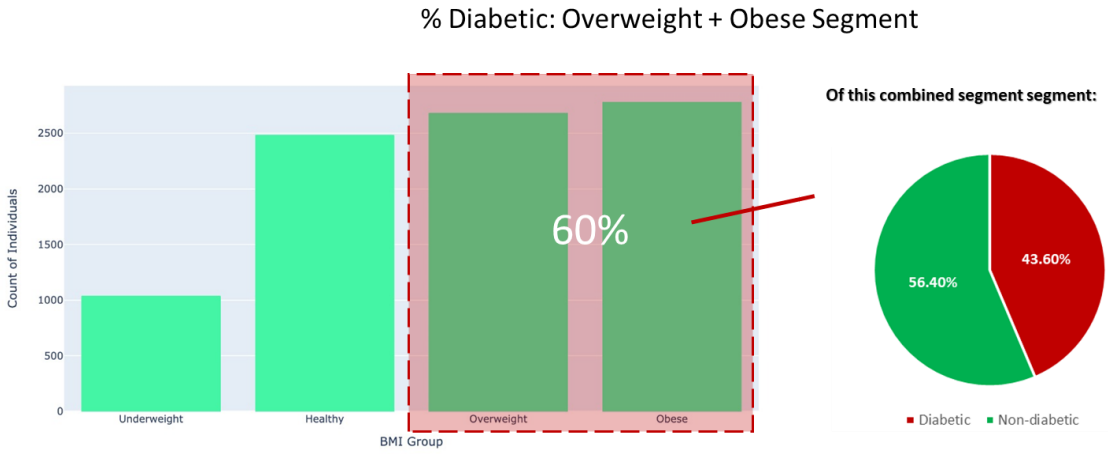




- In terms of race, whites and blacks show higher ratios of diabetics (*diabetics in race/ total race*) intragroup compared to other ethnicities. This may be owing to dietary preferences – an assumption that should be validated in a separate study.
- Likewise, higher income brackets show lesser proportion of diabetics intragroup compared to middle and low income classes. The specific reason cannot be intuited from the dataset but one can assume higher income affords more access to healthier, quality

food, less exposure of stress and more freedom to pursue exercise and other recreational activities – factors which reduce the risk of diabetes.

Section 3.2.2: A Medical EDA View of Diabetes



Analysis shows that:

- Easily 60% of the population is above the recommended BMI, with 31% of the population being obese and the remaining 30% classified as overweight.
- These sample figures echo statistics published by Harvard Research which estimates that 1 of every 3 US adults is obese[4] – an alarming statistic given the tight correlation between obesity and diabetes.

A side by side comparative of median BMI alongside average blood chemistry and nutrient values for diabetic and non-diabetic is equally informative.

- Higher levels of elements such as cholesterol and larger values for waist and arm circumference median of body index are indicators of poor health. Hence, said variables are potentially strong predictor candidates for whether a person has diabetes – a hypothesis the team seeks to confirm as part of the modeling process.

[4] *Adult Obesity | Obesity Prevention Source | Harvard TH Chan School of Public Health*. (n.d.). Harvard TH Chan School of Public Health. Retrieved November 21, 2021, from https://www.hsph.harvard.edu/obesity-prevention-source/obesity-trends/obesity-rates-worldwide/

DIABETIC

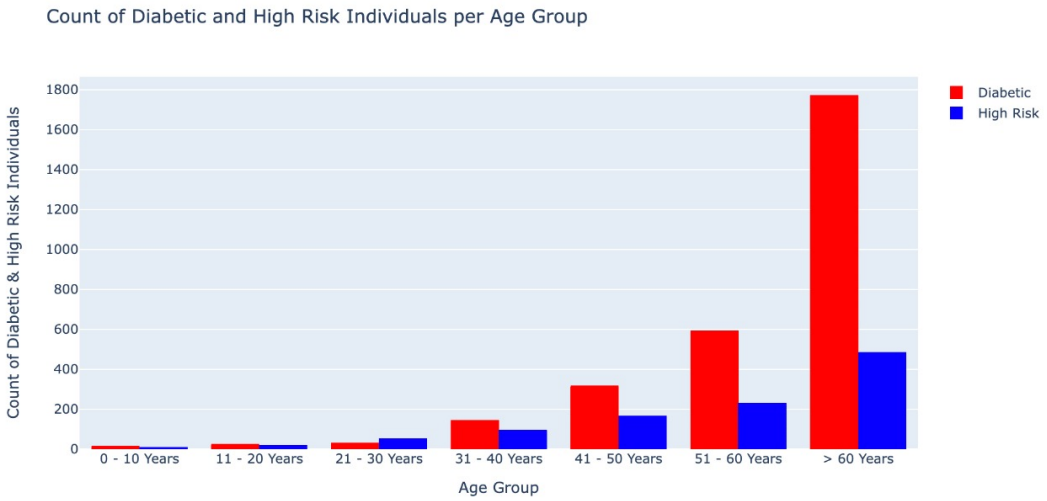
22.88 Average of Vitamin B6 (mg)	450.75 Average of Calcium (mg)	12.57 Average of Cholesterol (mg)	1.38 Average of Copper (mg)
0.67 Average of Cadmium (ug/L)	1.44 Average of Lead (ug/dL)	10.89 Average of Blood manganese (ug/L)	206.54 Average of Blood selenium(ug/L)
29.10 Median of Body Mass Index (kg/m**2)	103.83 Average of Waist Circumference (cm)	33.49 Average of Arm Circumference (cm)	5.85 Average of Glycohemoglobin (%)

NON-DIABETIC

18.83 Average of Vitamin B6 (mg)	423.96 Average of Calcium (mg)	12.51 Average of Cholesterol (mg)	1.37 Average of Copper (mg)
0.68 Average of Cadmium (ug/L)	1.38 Average of Lead (ug/dL)	11.32 Average of Blood manganese (ug/L)	206.37 Average of Blood selenium(ug/L)
25.30 Median of Body Mass Index (kg/m**2)	87.85 Average of Waist Circumference (cm)	29.27 Average of Arm Circumference (cm)	5.61 Average of Glycohemoglobin (%)

Plotting diabetic cases and high risk individuals across age groups, a clear trend emerges where:

- The number of positive cases and high risk increase as age increases.
- This is notably a steep curve for confirmed diabetics from age 31-40 to the >60 cluster
- A similar upward trend is observed for high risk individuals for the same age range.



A synthesis of all these exploratory findings informs the modelling objective which we articulate in the next section.

Section 4: Model Selection and Roadmap

Section 4.1: Model Objectives

To align with the goal of promoting customer health and reducing claims through early intervention., we define 2 key objectives for our model:

(a) Determine probabilities of being a confirmed diabetic for targeted intervention and
(b) Estimate expected claims to adjust policy premiums and deductibles depending on intervention results.

Section 4.1.1: Diabetes Probability Classification Model Rationale & Intent

Using the wealth of personal and medical information available, the ideal model is expected to address the first objective (a) by:

(a) computing the probability of diabetes given an individual's observed indicator values
(b) identifying good indicators for diabetes and measuring their predictive importance.

This will allow for targeted intervention to be undertaken, be that in the form of medical monitoring or referrals to lifestyle change companies.

To achieve the second objective (b), the model's confusion matrix can be overlaid with costs to calculate expected value of diabetes expenses claims. This can be used by insurance companies to reduce risk by proactively adjusting premiums based on intervention results.

Section 4.2: Modeling Roadmap

To execute on the above, the team followed a structured approach towards the model selection and build process.

Section 4.2.1: Data Pre-Processing and Modeling Considerations

Before running and training the model, a few pre-processing steps were undertaken.

- Features were derived or re-clustered from the original dataset as outlined in Section 3.1
- Standard scaler was applied to the entire data frame given the widely different scale of the data feature values.

PCA and Lasso (L1) regularisation were also explored to reduce the dimension of our data.

- PCA was attempted but later dropped in favor of model explainability since improvement to performance was only minimal.
- Instead, Lasso (L1) regularization is incorporated in hyper parameter tuning.

Lastly, we did a split for our training and test dataset.

- Data was divided into 70% train - 30% test segments using stratified train-test split. As positive cases only represented 33% of our data, a stratified methodology was chosen to counteract the imbalance of positive and negative cases.

Section 4.2.2: Model Choice and Hyper-parameter Tuning

Five curated algorithms were used to process the data, the details of which are shown below:

Model Name	Tuned with...	Hyperparameter Settings
SVC	• Grid Search CV	• C value =0.01
Logistic Regression	• Grid Search CV	• C value =0.1 • Penalty = L1
Decision Tree Classifier	• Grid Search CV	• Criteria = gini • Max depth = 8 • Min leaf samples =9 • Min samples split =6
Random Forest	• Randomized Search CV • Grid Search CV	• Criteria = gini • Max depth = 21 • Min leaf samples =2 • Min samples split =7
XGBoost	• Bayes Search CV • Randomized Search CV • Grid Search CV	• Objective = binary:logistic • Colsample_bytree = 0.73044 • Gamma =0.13031 • Learning rate = 0.04281 • Max depth= 13 • Min child weight =2 • N_estimators=68 • Subsample=0.74211

Section 5: Model Results & Feature Importance

Section 5.1: Model Performance

Each model was run using the tuned hyperparameters in Section 4.2.2 and evaluated across eight performance measures

Metric	SVC	LOGREG	DCT	RF	XGBOOST
Train accuracy	0.816	0.814	0.867	**0.987**	0.979
Test accuracy	0.791	0.798	0.798	0.821	**0.823**
TP rate	0.684	0.682	0.680	0.688	**0.697**
FP rate	0.156	0.144	0.143	**0.112**	0.114
TN rate	0.842	0.856	0.857	**0.888**	0.886
FN rate	0.315	0.318	0.320	0.313	**0.303**
Precision	0.686	0.702	0.704	0.754	**0.756**

AUC	0.76	0.77	0.77	**0.79**	**0.79**

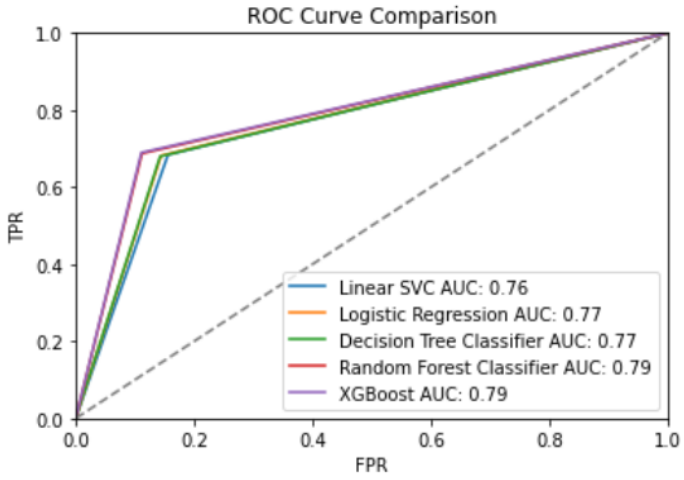
Taking into account class imbalance and the objective of identifying positive and high probability cases early on, there are 2 points to make when evaluating performance.

(1) **Accuracy is NOT the right metric:** This issue is evident in current methods which have decent accuracy but only because they predict most individuals to be non-diabetics when non-diabetics dominate in the population to begin with.

(2) The **confusion matrix is more instructive**. Based on our objective, the best model should have the below characteristics:

- **HIGH** sensitivity and specificity rates, AUC
- **LOW** fallout and miss rates

This can be seen in the below ROC plot and AUC value, which identify XGBoost to be the model of choice. While AUC values are the same with Random Forest, XGBoost has better true positive and false negative values - components which are important for evaluating claims EV in Section 6 of this paper.

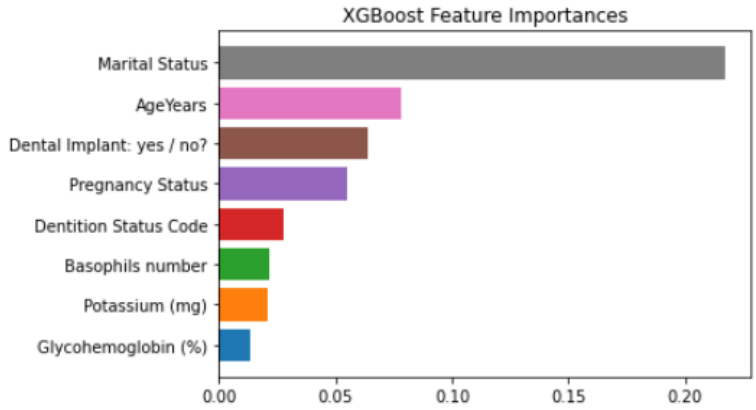


Section 5.2: Feature Importance

Feature importance analysis was done to understand (a) top indicators common across the 5 models and (b) those that heavily influence our best model (XGBoost). The below table summarizes the results of said analysis.

Metric	Common?	In XGBoost?	Commentary
Age	✔	✔	Older individuals are more likely to have diabetes

Glycohemoglobin	✔	✔	This indicates average blood sugar and is a standard test for diabetes.
Obesity indicators	✔		Obesity is highly correlated to diabetes. Measures of interest include average sagittal abdominal diameter, BMI, arm and waist circumference.
Oral Health Status	✔	✔	Indicators like whether an individual has implants are important as diabetics are more likely to develop oral health issues.[5]
Marital Status		✔	Marital status had a pronounced effect on a person's probability to have diabetes. This aligns with other correlational studies between marital status and other diseases like hypertension, for which there is a studied marked effect albeit different depending on gender[6].
Pregnancy Status		✔	While gender specific, pregnancy is often associated with gestational diabetes.
Potassium, Basophil count		✔	Potassium is associated with insulin production and therefore to sugar level while basophil count is usually elevated for people with type 2 diabetes.



Compared to traditional tests with only 30% sensitivity, the above features boost the company's predictive capability and puts it in a good position to realize the 2 business objectives initially defined. In the next section, the team talks through the business implication of this.

---

[5] Lamster, I.B., Lalla, E., Borgnakke, W.S., & Taylor, G.W. (2008). The relationship between oral health and diabetes mellitus. *Journal of the American Dental Association, 139 Suppl*, 19S-24S .

[6]Ramezankhani, A., Azizi, F., & Hadaegh, F. (2019). Associations of marital status with diabetes, hypertension, cardiovascular disease and all-cause mortality: A long term follow-up study. *PloS one*, *14*(4), e0215593. https://doi.org/10.1371/journal.pone.0215593

Section 6: Business Impact and Implementation

While model metrics are helpful, quantifying the benefit in dollar terms makes it easier to assess true business impact.

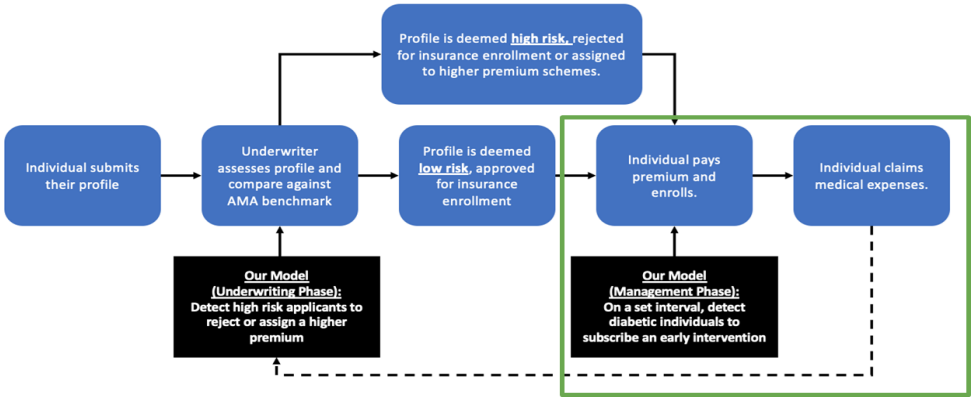
Section 6.1: Performance Comparison in Dollars

The team does this by using the 'Expected Claim (EC)' framework to assess three conditions

> **(1) Perfect Classifier**: This ideal model classifies all observations correctly. Plugging some values into the EC formula, the expected claims value would be **$5,575.70.** Performance will be judged based on the difference from this base value.
>
> **(2) Current Method**: The current standard yields an EC value of **$6075.37**.
>
> **(3) Our Model**: XGBoost run with tuned settings yields an EC value of $**5,778.13**.

Comparing the current standard with XGBoost shows a clear 59.49% uplift ($499.67 vs $202.43 differential per head). When superimposed on top of the 2,962 test individuals, this translates to over a $800,000 impact to the company bottom line.

Section 6.2: Business Implementation



With a clear uplift versus current standard, the team recommends implementing the model for existing customers. This has limited to no downside since the company will not be asked to turn down new customers nor will implementing this entail any change to operational flows.

Execution-wise, implementation would mean:

1. Feeding personal and medical data of existing customers into the XGBoost model.
2. Reviewing probability outputs and tagging those classified as diabetic for confirmation checks and premium reassessment.
3. Identifying high risk cases and tagging them for early intervention - be that through continuous medical monitoring or referring them to lifestyle change merchant partners.
4. Both subgroups would be continuously monitored and their results factored in the company's risk assessment models.

Viewed from a product lens, insurance companies can also use this as an opportunity for service differentiation. Rather than sticking to a purely risk hedging use case, it can use these new insights to :

1. Provide personalized insurance plans with pricing reflective of actual individual risk
2. Provide a partner app for overall health monitoring and needs matching to lifestyle change companies.

This allows it to diversify its revenue stream, pad the bottom line and build natural synergies with areas it clearly is connected to but has yet to explore. In Section 6.3, the team goes through a list of backlog items for the next iteration of the model and its subsequent application to other parts of the business process.

Section 6.3: Model Next Steps

Potential improvements and next steps in the design and implementation of the model could include:

1. Extend the model for usage during the underwriting phase for profile assessment and premium assignment. It can provide medical benchmarks for future profile evaluations.
2. Acquire more complete datasets with clearly defined scopes for features like pregnancy and marriage status. There are a high number of 'unknowns' in the current dataset.
3. Include labels for individuals who are already being treated for diabetes as some of their medical readings may differ from those who are untreated. This can be used as a control variable when required.
4. Gather data with features of individuals who are prediabetic and build a model for classification of prediabetics. Early intervention can mitigate the risk of them developing diabetes in the future.