PROGNOSTIC ANALYTICS

A Project Report

Submitted by

MARYAM RASHID

Roll No. 27

Towards the completion of degree of

Bachelors of Mathematics

CENTRE FOR ADVANCED STUDIES IN PURE AND APPLIED MATHEMATICS BAHAUDDIN ZAKARIYA UNIVERSITY MULTAN, PAKISTAN

Project Supervisor:

Dr. Athar Kharal

Abstract

This project focuses on predicting medical insurance costs using machine learning. The goal is to build a model that estimates patient charges based on factors such as age, BMI, smoking status, and region. Using Python libraries like Pandas for data handling and Scikit-learn for modeling, I implemented the Random Forest algorithm. The project covers data cleaning, exploratory analysis, and model evaluation to identify key cost drivers. The results show that smoking status and BMI have the strongest impact on insurance prices, and Random Forest provides more accurate predictions. The study demonstrates how machine learning can help insurers and patients better understand healthcare pricing.

Contents

Chapter 1

Introduction to Prognostic Analytics

1.1 Overview of Prognostic Analytics

Prognostic analytics is a branch of data analysis that focuses on understanding the causes and effects of events to predict future outcomes. Unlike predictive analytics, which basically forecasts future outcomes based on historical data, prognostic analytics delves deeper into the underlying factors that influence these outcomes.

1.2 Key Concepts in Prognostic Analytics

1.2.1 Definition

Prognostic analysis involves causal analysis to understand why certain events occur, enabling more informed predictions about future outcomes.

1.2.2 Significance and Relevance

Prognostic analytics plays a very important role in modern data-driven decision-making. Not only does it forecast future outcomes, but also uncovers the underlying causes behind those trends.
It plays a significant role in improved decision making, reducing risks, operational efficiency and AI & automation integration.
Its relevance covers industries-from healthcare to finance, and this is what makes it a foundational element of modern analytics.

1.3 Comparison with Predictive Analytics

- **Goal:** Predictive analytics looks at historical data, while Prognostic analytics focuses on cause and effect.
- **Applicability:** Predictive analytics gives probabilities or possible outcomes while Prognostic analytics recommend actions to influence results.
- **Applications:** Predictive analytics is applied in demand forecasting, customer churn, risk scoring, etc. Prognostic analytics has applications in predictive maintenance (with root-cause diagnosis), personalized treatment plans in healthcare, fraud prevention.
- **Example:** For example, in the case of equipment failure, Predictive analysis will give result as ”There is a 70% chance this machine will fail next month.” while Prognostic analysis will give answer as ” This machine will fail due to bearing wear; replace it now to avoid downtime.”.

1.4 Causal Analysis and Model Development Workflow

1.4.1 Data Collection

Data Collection is the process of gathering and measuring information on variables of interest through a well-defined and orderly process which allows one to answer questions and evaluate outcomes.

Sources of Data

- **Primary Data:** The data which is collected for the first time by the researcher, is Primary Data. It is collected with an aim for getting the solution for the probem at hand.
- **Secondary Data** is the data which is already collected or produced by others. For example, Government census reports.

1.4.2 Data Pre-processing

Data Pre-processing includes the techniques and procedures that are used to prepare raw data in a clean, organized and structured format which is suitable for analysis or modeling.

Benefits

- Improved Data Quality
- Enhanced Model Performance
- Reduced Computational Complexity
- Ensures Compatibility

1.4.3 Feature Engineering

Feature Engineering is a process in which we transform raw data into meaningful data(features) that helps the models to make better predictions. It improves the accuracy of model. The steps in feature engineering involves selection, transformation. It basically relies on domain knowledge for validating the data.

1.4.4 Causal Analysis

The type of analysis in which main thing of study is if the change in one variable cause change in another variable in dataset. It identifies causes and effects of a phenomenon, problem or event. It helps in answering certain questions like:

- Why did something happen?
- What are the consequences of happeing of something?
- How can something be prevented/improved?
- What are best solutions?

Direct Causes

An event that immediately and necessarily produces an effect without any mid-level steps is called Direct Cause. Prognostic analysis progresses on the direct causality by turning the raw data into actionable and accurate predictions. Without the concept of direct causes, only correlations are left that may not help prevent breakdowns.

Indirect Causes

The underlying or hidden factors which become a cause of a failure but do not directly trigger it, are Indirect causes.They create conditions for the direct cause to occur.
The actual hidden enablers of failure are indirect causes. As prognostic models prioritize direct causes for real-time predictions, incorporating indirect causes may ensure comparatively most effective solutions.

1.4.5 Model Development

Model Development is actually the process of creating, training and enhancing the machine learning model to obtain meaningful results from data and get solutions of complex problems. The main part of model development is training and tuning the model. It involves splitting the data into training and testing sets.
Some benefits of model development are:

- Enhanced Decision Making
- Early Risk Detection
- Scalability

- Competitive Advantage
- Effeciency and Automation

1.4.6 Model Validation

Model Validation is the set of processes that are designed to ensure whether a model is performing the way it should be. It includes both its design goals and how it benefits the end user. An important part of validation is testing whether a model is working correctly in a real world setting or not.

In prognostic analytics, this process ensures that a model accurately predicts failures and identifies the root causes, ensuring generalization and fairness.

Model Validation matters due to following reasons:

- Validation uncovers weaknesses that help to refine the model over time.
- Validated models are more trustworthy.
- They can help prevent costly mistakes.

1.4.7 Deployment and Monitoring

Model Deployment means making the model available to users or systems so that they can make decisions based on data, interact with their application, and so on. Model Monitoring is the process of closely monitoring the performance of models. It provides with the conclusions of performance and health of the deployed model. Without deployment, models are only experiments, without any solutions. Similarly, monitoring is a crucial process for keeping a check of the model's performance.

1.4.8 Generating Insights

Generating insights is a critical step in ensuring that the model generates meaningful results. Insights are the implementable evidence-based discoveries that uncover:

- Why does something happen? (Root Causes)
- When they are likely to happen? (Predictive)
- How to prevent them? (Prescriptive)

1.4.9 Strategic Actions

After deriving the insights from a prognostic model, the next step is to perform applicable strategies. It involves converting these findings into a definite plan. It helps to make informed decisions to obtain impactful results.

1.5 Case Study: Prognostic Analytics in Medical Insurance Pricing

1.5.1 Problem Statement

Many people struggle with high medical insurance costs but have trouble understanding what makes their premiums expensive. The Insurance companies rely on complex models to determine pricing that are not easily understandable by companies. Similarly, insurers struggle with fair pricing with financial sustainability. This project uses prognostic analytics to analyze dual perspective, and develop a fair pricing framework.

1.5.2 Methodology

Approach

Used prognostic analytics to identify causal relationships and use individual profiles to anticipate future costs.

Tools

Python libraries like NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, and Random Forest algorithm are used.

Data

- **Dataset:** The dataset consists of 1338 policy-holders.
- **Features:** Features include age, BMI, sex, smoking status, and region.
- **Pre-processing:** The categorical variables are encoded into numerics (smoker → 1/0, region → 0-3).

1.5.4 Conclusion

This case study highlights how machine learning is reshaping premium pricing through smarter, data-driven decisions. Our predictive model accurately forecasts the costs for new customers 85% of the time, typically within $2,700 of their actual expenses. It also identified key risk factors, most notably smoking, which contributes to 61.5% of the risk, along with BMI and age. These insights allow for fairer pricing and open the door to meaningful cost-saving strategies.

Dual-Perspective Analysis

1. Insurer's Perspective
Following are the key findings:

- The model reduces pricing errors by 15% (85% accuracy) vs. traditional methods.
- Smokers cost 4× more to insure.

These strategic actions can be taken for insurers:

- Wellness programs targeting smoking could save $1.98M/year per 100 policyholders.
- Adjust premiums annually using updated health data (e.g., BMI improvements).
- Offer discounts for healthy behaviors (e.g., gym memberships, smoking cessation programs).

2. Policy Holder's Perspective
Following are the key findings:

- SHAP values show exactly why premiums increase.
- Smoking adds $15,000 to your cost

These strategic actions can be taken by policyholders:

- Quitting smoking could lower premiums by 80%.
- Annual health check-ups to prove reduced risk like lower BMI.

Future Outlook

By integrating predictive analytics with active policyholder engagement, insurers can move beyond traditional pricing models toward a more sustainable and customer-centric approach.
This strategy enables fairer, behavior-based premiums that reward healthy choices while improving risk management.
Rather than just forecasting costs, the model helps reduce them, resulting in fostering a healthier, more transparent insurance ecosystem that benefits both insurers and policyholders.