# Beyond Accuracy: A Mock AI Audit for Fair and Transparent Insurance Cost Prediction with Low Code and No Code.

Sushika Reddy Gade,

Faculty Advisor: Arup Das

School of Sciences, Monmouth University, West Long Branch, NJ

**Abstract**

The increasing adoption of machine learning (ML) in insurance pricing raises critical concerns regarding **fairness, transparency, and regulatory compliance**. Recent regulations, such as the **New York City Automated Employment Decision Tool (AEDT) Law** and **Colorado’s AI Insurance Pricing Regulation**, highlight the need for systematic AI audits to ensure fairness and mitigate bias in high-stakes decision-making.

This study conducts a **mock AI audit** of an XGBoost-based ML model for medical insurance cost prediction, going beyond traditional accuracy metrics such as **R² score and RMSE**. Instead, we emphasize **fairness, bias detection, and model explainability**, employing **SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations)** to assess feature influence—such as **smoking status, age, and BMI**—and detect biases. If bias is identified, we explore **mitigation strategies**, including **fairness constraints and feature adjustments**, to enhance equitable decision-making.

A key innovation of this study is its focus on **democratizing AI auditing** through **low-code and no-code platforms**, making fairness assessments accessible to **business analysts, regulators, and decision-makers without coding expertise**. These platforms provide intuitive visual tools, enabling broader stakeholder participation in AI governance and regulatory compliance.

Our findings contribute to a **best-practices framework** for AI audits, emphasizing that ML models must be evaluated **not only for accuracy but also for their societal and ethical impact**. By integrating **regulatory compliance, fairness assessments, and explainability techniques** within a **low-code/no-code framework**, this study presents a **novel and practical approach** to ensuring **transparent, fair, and accountable AI systems** in insurance pricing and beyond.

**Keywords**

Health Insurance pricing, Responsible Artificial Intelligence, Machine learning, Insurance claims, Healthcare, AI audit, no code/low code.

**Introduction**

Health insurance plays a key role in an individual’s life by financially safeguarding against unhinged healthcare prices in the United States of America. Processing insurance claims accounts for a significant amount of healthcare costs, making the US healthcare system one of the biggest and most intricate in the world. There has been a significant increase in integrating machine learning models to improve efficiency, reduce costs, and also enhance customer experience. This has changed how insurers assess risk, evaluate claims, and set premiums. It is evident that now predictive models play a significant role in giving an estimate for insurance costs, considering several customer demographics, including health indicators, and also factors that relate to lifestyle. However, as AI/ML gets increasingly ingrained in such decision-making, it also raises concerns with responsibility, transparency, fairness, and accountability.

A sense of urgency for this point at issue is enlarged by the increasing costs of healthcare in the United States. According to National Health Expenditure (NHE) projections, healthcare spending is expected to grow at an average annual rate of 5.6% from 2023 to 2032, outpacing the projected GDP growth of 4.3%. This will increase healthcare’s share of the economy from 17.3% of GDP in 2022 to 19.7% in 2032. The expansion of insured populations, including a record 91.2 million Medicaid enrollees in 2023, has driven greater healthcare utilization, pushing expenditures higher. Although Medicaid enrollment is expected to lower after 2024 due to policy changes, rising healthcare prices, increased service utilization, and prescription drug cost drift under the Inflation Reduction Act (IRA) will continue to strike up insurance costs. This evolving area presents a challenge for insurers as they strive to stabilize affordability, fairness, and profitability in premium pricing.[[1]](#bookmark=id.atcaz2suus2f)

One of the concerning factors is AI’s pricing models. It might worsen already existing healthcare cost injustices while increasing efficiency and accuracy. It is well known that machine learning models are trained on historical data, and if the data persists of biases such as disparities in access to care or socioeconomic inequalities, the model may reinforce these biases in insurance pricing[[2]](#bookmark=id.dn4eg9945t1). Features such as age, smoking status, body mass index (BMI), and pre-existing conditions heavily influence premium calculations, and without proper oversight, pricing models may improperly penalize vulnerable populations. As a result, regulators, legislators, and industry participants are calling for AI models that are open, understandable, and consistent with moral standards.

While AI-driven insurance models provide commendable predictive capabilities, most of them work as black-box models. Black-box models are complicated machine learning algorithms, such as neural networks and ensemble learning, where the flow of their internal processes is not intelligible and interpretable by humans. This leads to a lack of transparency, raising critical concerns for regulators, insurers, and also customers. It becomes extremely difficult to determine why a particular individual receives higher premium quotes or why a particular feature has a significant impact on determining costs.

However, without proper explanation, the faith in AI-driven insurance pricing can tarnish as potential biases remain unaddressed. An insurer can unknowingly develop an insurance pricing model without noticing the underlying biases in the model. This can lead to disproportionately raising the premiums for specific demographics without actually knowing the reason. To tackle this issue, researchers and policymakers have emphasized explainable AI (XAI) methodologies [[3]](#bookmark=id.ivbv7anv9jmf). These methodologies help us in interpreting AI-driven decisions, identifying underlying biases, and also enhancing the fairness of the model.

**Regulatory Response: The Role of AEDT in AI Governance**

To make sure that decisions that are made by AI are fair, accountable, and explainable, regulatory frameworks have started addressing the ethical concerns around AI’s decision-making. An example is **New York City’s Automated Employment Decision Tool (AEDT)** law [[4]](#bookmark=id.f1i8ioahu0r5), which came into effect in July 2023. This law mandates audits for AI-driven employment decisions to detect and mitigate bias in hiring and promotion processes. While AEDT primarily focuses on employment decisions, it indicates a broader trend of AI regulation that could extend to sectors like health insurance, loan approvals, and credit scoring, where AI significantly influences financial outcomes for individuals.

In addition to AEDT, **Colorado’s AI Insurance Pricing Regulation** represents a landmark step in AI governance for the insurance industry[[5]](#bookmark=id.jxn8svuf96f8) . Colorado became the first U.S. state to pass a regulation that directly governs the use of AI in insurance pricing, requiring insurers to demonstrate that their AI models do not unfairly discriminate against protected classes. The law mandates insurers to conduct bias audits and provide transparency reports on AI decision-making processes. This regulation reflects a growing recognition of the need for fairness and accountability in AI-driven pricing, ensuring that automated systems do not reinforce systemic biases that could disproportionately affect vulnerable populations.

**Mock AI Auditing and No-Code/Low-Code Solutions**

Similar to the above-mentioned regulatory efforts, this study conducts a mock AI audit of an ML model designed for medical insurance cost prediction. We explore how low-code/no-code AI platforms can democratize AI auditing. Generally, auditing AI models requires in-depth technical expertise, confining the ability specifically to data scientists. We leverage **Orange,** a visual programming tool for data science and machine learning**.** Orange is an open-source no-code/low-code platform that enables users to build, analyze, and interpret machine learning models through an interactive graphical interface [[6]](#bookmark=id.a09q8r833u94)**.** By using Orange, we mitigate the need for complex programming, allowing non-technical stakeholders such as insurance regulators, policymakers, and business analysts to engage with AI audits.

This simulated audit critically examines the model's decision-making process, interpretability, and ethical risks. We have used Gratel.AI[[7]](#bookmark=id.w3jt5kpdktbz) to create a synthetic insurance pricing dataset and evaluate the model’s predictive performance using R² and RMSE. In addition to that, we emphasize fairness, bias detection, and model explainability, employing SHAP (Shapley Additive Explanations)[[8]](#bookmark=id.4znzuqit7lsj),[[9]](#bookmark=id.5ok6bor78m1y) to assess feature influence and investigate whether the model treats all policyholders impartially.

Additionally, by offering insights into best practices for AI auditing, bias detection, and transparency in healthcare pricing models, our study adds to the larger conversation on AI fairness. We highlight the significance of easily accessible AI auditing tools that enable wider stakeholder interaction in order to ensure justice and accountability in AI-driven decision-making by utilizing low-code/no-code solutions such as Orange.

**Insights about our Dataset**

For this study, we used a dataset designed to reflect the kinds of information insurers often rely on to determine medical insurance costs. Each row in the dataset represents an individual applicant, and each column contains details about that person's background, lifestyle, health status, or financial profile. The final column, called Price, is the target variable; this is the predicted insurance premium for each person.

The dataset includes a total of 18 variables, which we grouped into two types: numerical and categorical.

The numerical features are things that can be measured or counted, like:

1. Age
2. Income
3. BMI (Body Mass Index)
4. Deductible
5. Prior\_Claims
6. State\_Insurance\_Cost\_Index
7. Inflation\_Adjusted\_Factor
8. Credit\_Score
9. Price (the outcome we’re predicting).

These variables help capture measurable factors like how old someone is, how much they earn, or how many insurance claims they’ve made in the past.

The categorical features, on the other hand, describe traits or classifications. These include:

1. Gender
2. State
3. Employment\_Status
4. Smoking\_Status
5. Pre\_Existing\_Conditions
6. Exercise\_Frequency
7. Plan\_Type
8. Coverage\_Level
9. and Education\_Level.

**Data Dictionary**

| **Column Name** | **Description** |
| --- | --- |
| Age | Age of the insured individual (18-80) |
| Gender | Gender of the individual (Male/Female) |
| State | U.S. state where the individual resides |
| Income | Annual income in USD (20,000 - 150,000) |
| Employment\_Status | Employment status (Employed, Unemployed, Self-Employed, Retired) |
| BMI | Body Mass Index (18.5 - 40) |
| Smoking\_Status | Whether the individual is a smoker (Smoker/Non-Smoker) |
| Pre\_Existing\_Conditions | Medical conditions such as Diabetes, Hypertension, Heart Disease, etc. |
| Exercise\_Frequency | Frequency of physical exercise (Never, Rarely, Sometimes, Often, Daily) |
| Plan\_Type | Type of insurance plan (Basic, Standard, Premium) |
| Deductible | Deductible amount in USD (500 - 5000) |
| Coverage\_Level | Level of insurance coverage (Low, Medium, High) |
| Prior\_Claims | Number of past insurance claims (0-4) |
| State\_Insurance\_Cost\_Index | Multiplicative cost factor based on state (0.8 - 1.2) |
| Inflation\_Adjusted\_Factor | Adjustment factor based on economic conditions (0.9 - 1.1) |
| Credit\_Score | Individual credit score (300 - 850) |
| Education\_Level | Highest education level attained (High School, Associate, Bachelor, Master, PhD) |
| Price | Predicted insurance cost in USD (500 - 20,000) **(Target Variable)** |

**Table 1: Description of synthetic health insurance dataset**

**Key Features:**

**Demographic Attributes:** Age, Gender, State, Income, Employment Status

**Health & Lifestyle Factors:** BMI, Smoking Status, Pre-Existing Conditions, Exercise Frequency

**Insurance Details:** Plan Type, Deductible, Coverage Level, Prior Claims

**Economic & Geographic Indicators:** State-Based Insurance Cost Index, Inflation Adjusted Factor

**Bias-Sensitive Attributes:** Credit Score, Education Level

**Target Variable:** Price (Insurance Cost Prediction)

Each of these categories gives context to an individual's lifestyle or choices, like whether they smoke, how often they exercise, or what type of insurance plan they’re enrolled in. These are all important to understanding what drives insurance costs.

We also included some external factors like regional cost adjustments (State\_Insurance\_Cost\_Index) and inflation-related metrics (Inflation\_Adjusted\_Factor). These help account for broader economic conditions that might influence premiums, even if the person’s profile doesn’t change.

Together, these features in [Table 1](#bookmark=id.9ymgq4uepzkn) create a rich, realistic dataset that lets us explore how different variables, both personal and contextual, contribute to the way insurance prices are predicted. Having a well-defined and thoughtful data structure like this is crucial for building trustworthy models and evaluating them fairly, especially when we’re trying to audit those models for bias or explainability.

**Material and Methods**

**What is Orange?**

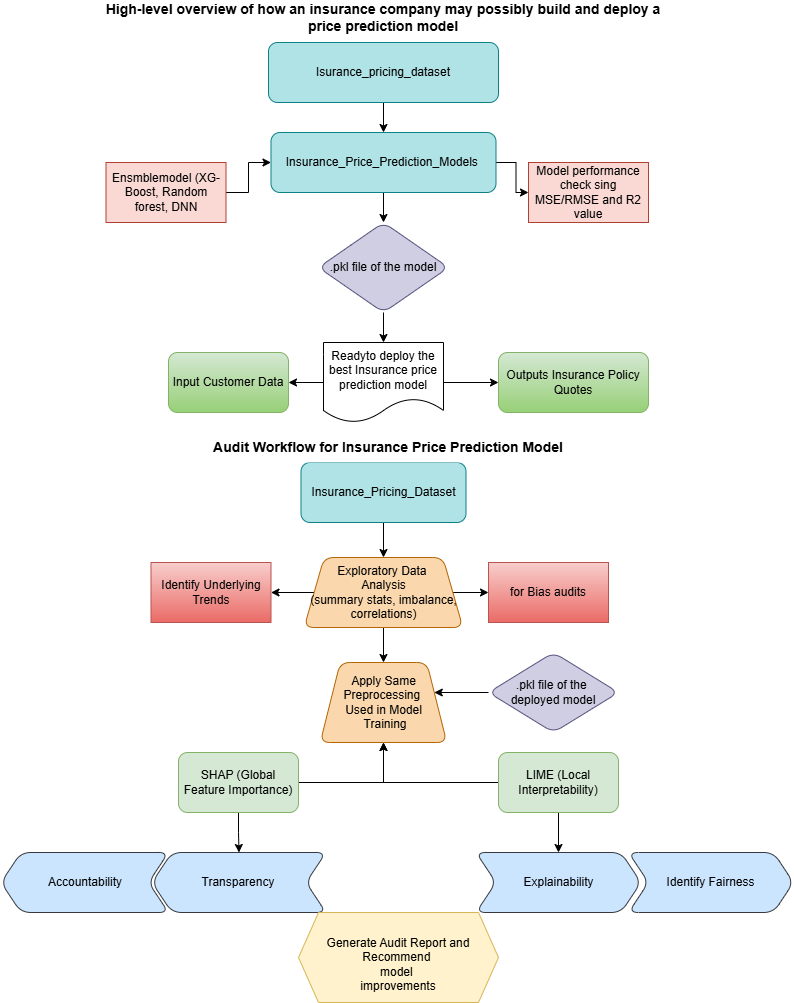
Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis. Orange components are called widgets, and they range from simple data visualization, subset selection, and preprocessing to empirical evaluation of learning algorithms and predictive modeling. We place the widgets on a drawing board (the “canvas”). Widgets communicate by sending information along with a communication channel. An output from one widget is used as input to another. Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration. Orange helped us design and analyze our workflow using an interactive interface, without the need for extensive programming.

**Why Orange?**

To carry out our analysis and model auditing, we used a tool called Orange, a no-code, visual platform that played a central role for us to work with machine learning without writing lots of code. Orange lets users drag and drop components to build data workflows, which was especially helpful for us as this made our project more accessible and gave us the flexibility to focus on ethical evaluation and explainability rather than just coding.[[6]](#bookmark=id.a09q8r833u94)

One of the biggest reasons we chose Orange is because it’s interactive and beginner-friendly. It helped us explore our dataset, choose the most important features, and evaluate how well our model was performing, all through a visual interface. We used Orange to load and explore our dataset, apply data preprocessing steps, connect a pre-trained machine learning model, and evaluate its performance using regression metrics like R² and RMSE, all within a single, visual environment.Orange’s no-code approach was especially valuable in supporting the ethical goals of our project by making the modeling and auditing process more accessible, explainable, and inclusive for those without a programming background.[[6]](#bookmark=id.a09q8r833u94)

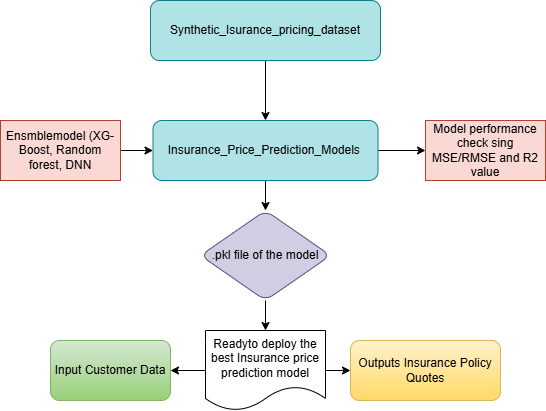
**Approach**

****

**Figure 1**

The mock AI audit approach we followed for evaluating a machine learning model built for insurance price prediction **is** divided into two key phases: Model Development (how an insurance company might build a model) and Audit Workflow (how an auditor would examine that model for ethical compliance and transparency) as illustrated in [Figure 1](#bookmark=id.hd3ifrb7pmzg)

### **Phase 1: Hypothetical Model Development Scenario by an Insurer**



**Figure 2**

This phase (as in [Figure 2](#bookmark=id.wug7kora171o)) represents how an **insurance company** might develop a machine learning model to predict insurance pricing based on customer data. It mimics a real-world deployment environment where predictive modeling is part of the core business decision-making process.

### **Start with a Synthetic Insurance Pricing Dataset and EDA.**

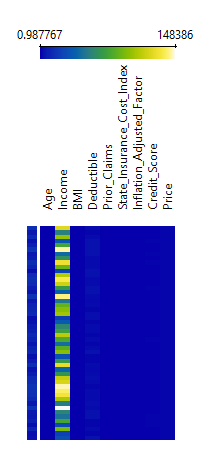
We used a synthetic dataset that mimics real insurance data. This dataset serves as the foundation for both training the model and conducting ethical audits later. We perform Exploratory Data Analysis (EDA) to understand feature distributions, detect missing values, explore relationships (e.g., smoking and BMI vs. cost), and spot any data imbalance. EDA helps inform preprocessing decisions and model development strategies.

We built the model in Orange using its components/widgets. Our orange workflow starts with the file widget. The File widget [reads the input data file](https://docs.biolab.si/3/visual-programming/loading-your-data/index.html) (data table with data instances) and sends the dataset to its output channel. The history of the most recently opened files is maintained in the widget. The widget also includes a directory with sample datasets that come pre-installed with Orange. In the schema, the widget is used to read the data that is used for training the model. The Data Table widget previews the attribute values in a spreadsheet. In this component, data instances may be sorted by attribute values. The widget also supports manual selection of data instances. Color widget adds visual color grouping to enhance interpretability in downstream visual widgets. For example, we can color data points by the ‘Price’ variable or ‘Smoking\_Status’ or any other instance to see their influence in our visualizations. The ‘Data info’ widget is used to get the technical summary of the dataset. It gives information on column names, data types, and missing values, if any. This widget is essential for cleaning the data before we head to modeling.



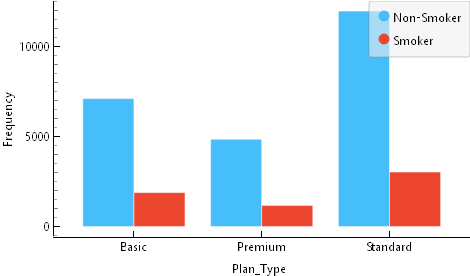
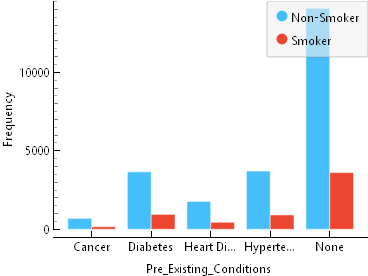
**Figure 3**

In order to get summary statistics like mean, median, mode, min/max, and standard deviation, we used the ‘Feature Statistics’ widget. This widget is useful in inspecting potential data entry errors and also skewed distributions. It is a quick way to inspect and find interesting features in a given data set, as shown in [**Figure 3**](#bookmark=id.hfr5kkscq42s), without any need to code. Then we used the ‘Correlations’ widget that computes correlations between numerical variables like ‘Price,’ ‘Age,’ or ‘BMI’ to identify linear relationships. We connected the correlations widget to a ‘heatmap’ widget to visualize ([Figure 4](#bookmark=id.e5co5ovknbu1)) the correlation matrix using color gradients.



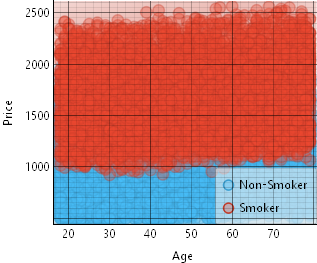
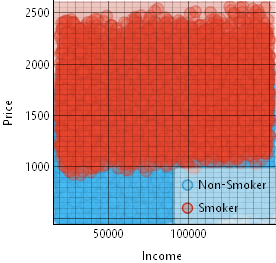
**Figure 4**

For the EDA, we have used a component of Orange called the ‘Distributions’ widget. It shows histogram bar s of each feature so you can detect skewness, understand how frequently values occur, and spot class imbalance in categorical variables like ‘Gender,’ ‘Plan\_Type,’ etc. [Figure 5](#bookmark=id.9we7mc8xonv) presents a bar chart, illustrating how insurance-related features like ‘Plan\_type,’, etc., are distributed across smoking status categories (smoker and non-smoker). And a bar chart in [Figure 6](#bookmark=id.9we7mc8xonv) illustrates that non-smokers dominate across all condition categories, especially in the "None" category, indicating a higher concentration of healthier individuals who do not smoke.

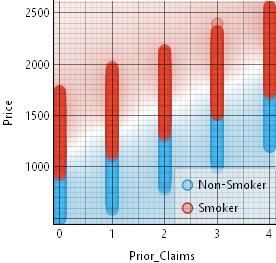
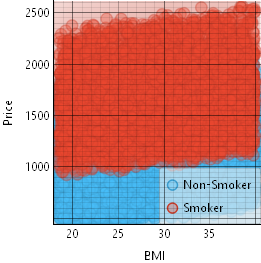
**Figure 5 (distribution of Plan\_Type for smoking status)** **Figure 6(Distribution of pre-existing conditions for smoking status)**

A scatter plot is used to visualize two numeric variables at a time to identify relationships or clusters, as shown in [Figure 6](#bookmark=id.9we7mc8xonv), which visualizes the relationship between ‘Age’ and ‘Price’ while differentiating between Smoker and Non-Smoker categories using color-coded data points. It is important for all relevant features to be visualized by the insurers to check if the model disproportionately penalizes smokers (or any demographic group in the dataset) without adjusting for other health indicators (e.g., BMI, existing conditions); it could lead to unfair pricing.

**Figure 7** **Figure 8**

**(scatter plot for age vs price filtering smoking status) (scatter plot for income vs price filtering smoking statu)**

**Figure 9** **Figure 10**

**(scatter plot for prior claims vs price filtering smoking status) (scatter plot for BMI vs price filtering smoking status )**

From Figures [7](#bookmark=id.4n54atly5edf), [8](#bookmark=id.y8vgdjvha3qm), [9,](#bookmark=id.uawb9epsqqg8) [10](#bookmark=id.uawb9epsqqg8) we can note that smokers consistently receive higher insurance pricing than non-smokers across variables like age, income, prior claims, and BMI. The red regions representing smokers are concentrated at higher price ranges, indicating that smoking status strongly influences premium costs regardless of other attributes.

The Exploratory Data Analysis (EDA) we performed before preprocessing helps in data understanding before model building and reveals any hidden patterns in the dataset. It also ensures the dataset is clean, complete, and representative. This particular step builds a solid foundation for having an accurate model.

**Preprocessing the Dataset**

The ‘Impute’ widget is used to handle missing values in the dataset. In the context of insurance data, missing entries (e.g., for income, credit score, or medical history) can distort model training and predictions. By replacing these missing values with appropriate substitutes (such as mean, median, or mode), we ensure that the model isn't biased or misled due to incomplete data. For insurers, this reflects a real-world scenario where customer data might be incomplete or partially disclosed.

The ‘Preprocess’ widget applies multiple data transformation steps, such as normalization, discretization, or feature scaling. These steps help standardize data formats and bring numerical features to a similar scale, which is essential for algorithms like neural networks and XGBoost to perform optimally. For insurance companies, preprocessing ensures that features like age, income, and BMI, despite having different units and scales, are fairly treated by the model.

The ‘Rank’ widget is used to evaluate and rank the importance of features based on their contribution to the target variable (in this case, insurance price). It uses statistical measures like information gain or the Gini index. For an insurer, this helps determine which factors (e.g., smoking status, BMI, or prior claims) are most influential in pricing, supporting transparency and regulatory compliance

The Select Columns widget helps choose the specific input (feature) and target (label) variables to be used in model training. It can also exclude irrelevant or redundant features from the dataset. Insurers would use this to define which customer attributes are used to predict insurance prices, and to exclude identifiers or sensitive features not meant for prediction

The Data Sampler widget allows us to split the dataset into training and testing portions or reduce the dataset size for quick experimentation. We typically use it to randomly select a subset of data for training and reserve the rest for testing. For insurers, this mimics real-world model validation, ensuring the model is trained on one portion of data and evaluated on unseen samples to measure generalization.

Use of these widgets in the workflow reflects the responsible preparation of data in a real-world insurance setting. They ensure the data used for training is clean, relevant, and fair, laying the groundwork for building a robust and interpretable insurance price prediction model.

### **Train Models Using Ensemble and Deep Learning Techniques**

We focused on building multiple models using advanced machine learning algorithms. This approach simulates how an insurance company might experiment with different modeling techniques to identify the most accurate and reliable method for predicting insurance prices. Below is a breakdown of the **Orange widgets** used during this phase.

The **Random Forest** widget implements an ensemble of decision trees trained on various subsets of the data. Each tree is built from a random sample of the data and a random subset of features. It reduces overfitting and improves generalization. It helps detect which features (e.g., smoking status, BMI, prior claims) are most influential in price predictions and can handle complex interactions in data.

The **XGBoost** widget is a powerful gradient boosting implementation that builds trees sequentially, each one improving the errors of the previous. It is known for its high predictive accuracy and ability to manage missing values and outliers. It allows insurers to model non-linear relationships, especially in large and complex datasets, making it ideal for pricing policies with many influencing factors.

The **Neural Network** widget represents a deep learning model capable of capturing non-linear patterns and interactions between variables. It can model complex relationships and capture subtle variations that simpler models might miss. It helps discover deep patterns (e.g., how combinations of lifestyle and health factors affect pricing) that are not obvious through traditional models.

This step mirrors an insurer’s model experimentation phase, where different machine learning techniques are tested with proper preprocessing and evaluation. The use of ensemble methods like XGBoost and Random Forest, alongside Deep Neural Networks, allows for a thorough exploration of modeling options to find the best performing algorithm for accurate and fair insurance pricing.

### **Evaluate Model Performance**

Each model was evaluated using key regression metrics as shown in [table 2](#bookmark=id.jiaa9coko2il):

1. MSE (Mean Squared Error): Measures the average squared difference between predicted and actual values.
2. RMSE (Root Mean Squared Error): Gives more weight to large errors, easier to interpret.
3. R² (Coefficient of Determination): Shows how well the model explains the variation in price.

Test and Score widget is used to evaluate and compare the performance of all trained models using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² as in [table 2](#bookmark=id.jiaa9coko2il). It Helps determine which model offers the best balance between accuracy and generalization.

| Model | MSE | RMSE | MAE | MAPE | R² |
| --- | --- | --- | --- | --- | --- |
| | Random Forest | | --- | | | 14700 | | --- | | | 121.25 | | --- | | | 95.020 | | --- | | | 0.081 | | --- |  |  | | --- | | 0.905 |
| | XG-Boost | | --- | | | 253.50 | | --- | | | 15.925 | | --- | | | 12.583 | | --- | | | 0.010 | | --- |  |  | | --- | | 0.998 |
| | DNN | | --- | | | 647.38 | | --- | | | 25.442 | | --- | | | 20.164 | | --- | | | 0.017 | | --- | | 0.996 |

**Table 2**

The best model based on high R² and low RMSE from [Table 2](#bookmark=id.jiaa9coko2il) was XG-Boost, which we selected for deployment.

Once the best model was finalized, it was saved as a .pkl file using Orange’s ‘save model’ module. This step reflects how a deployed model might be stored and reused in production settings.

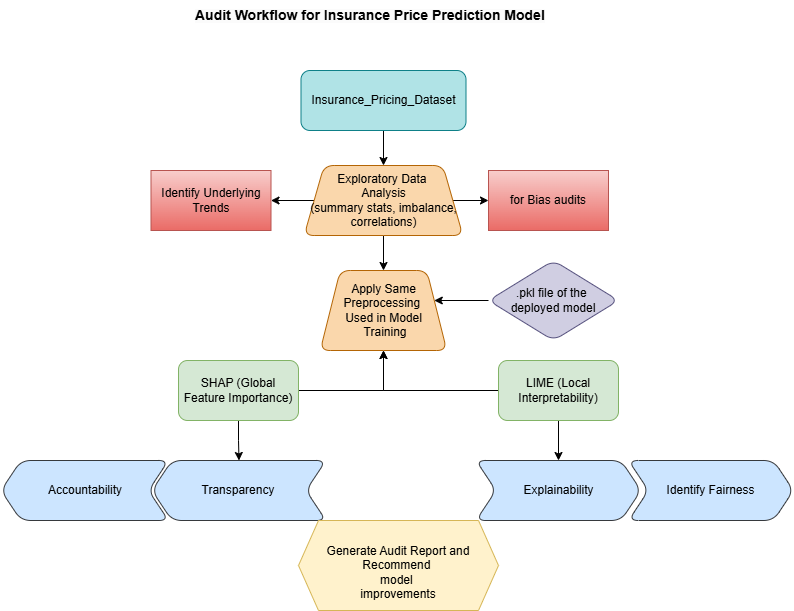
**Audit methodology**

### **Phase 2: Audit Workflow for Insurance Price Prediction Model**

Once the model is deployed, it enters a new phase of scrutiny. This phase simulates how regulators, auditors, or internal ethics teams might assess the fairness and transparency of the model.

**Load the original dataset for analysis.**

The audit begins by reloading the original dataset using orange’s ‘File’ widget. This step ensures the audit is done using the original input data to maintain consistency. The goal is to understand the structure of the data, identify sensitive attributes (e.g., gender, BMI, smoking status), and, explore potential biases.

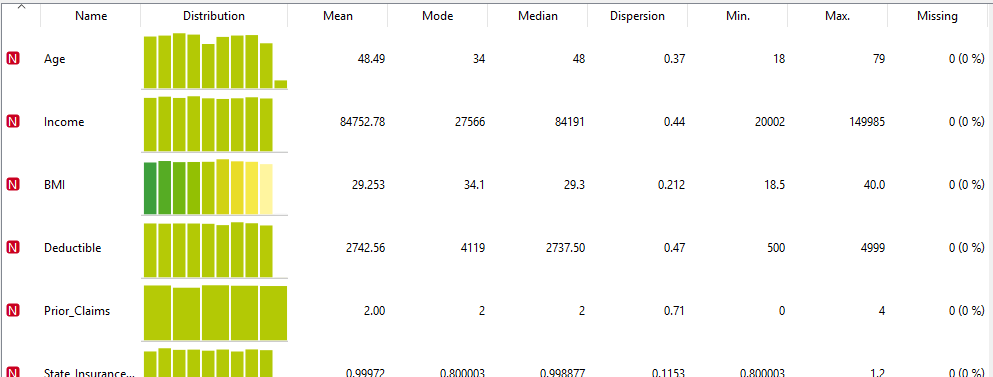


**Figure 11**

### 

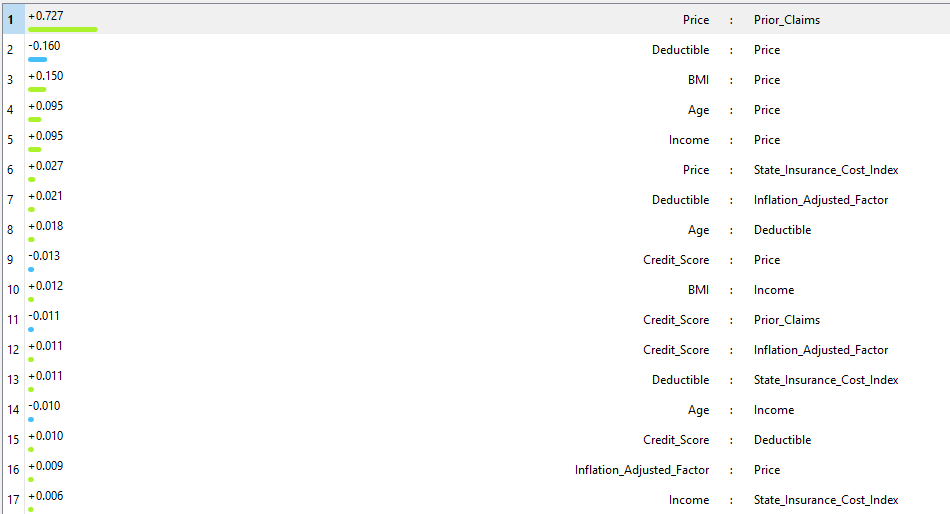
### **Exploratory Data Analysis (EDA)**

EDA is a key step in the audit process as illustrated in [figure 11](#bookmark=id.puiz71kv1icj). We visually inspect the raw dataset using various orange components like ‘data table,’ it previews the attribute values in a spreadsheet, which is helpful to get a basic feel for the structure, missing values, and outliers.. Using color widget, we can color and filter data by key features like "Smoker" or "Gender" as done earlier. ‘Data info’ widget is used to get a a quick overview of the dataset: number of rows, columns, missing values, and data types. This is essential for verifying dataset completeness and cleanliness before analysis or modeling. ‘Feature Statistics’ widget helps auditors understand the central tendencies and spread of numerical features like ‘Age’,’ Income’,’ BMI’, and ‘Credit Score’ ([Figure 12).](#bookmark=id.xxaqjsyfnwae) This step can potentially reveal anomalies, such as extremely high or low values that may influence model predictions.



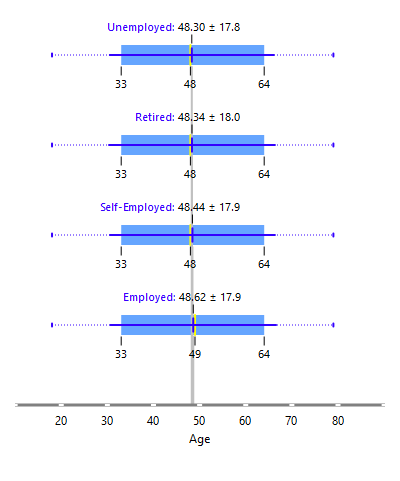
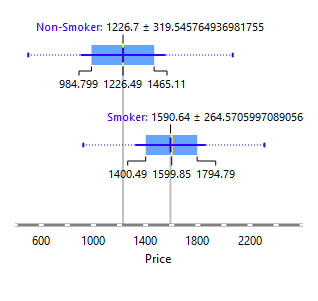
**Figure 12**

Using the ‘Correlations’ widget, we displayed pairwise correlation coefficients (like Pearson) as shown in [Figure 13](#bookmark=id.doc9q5n1n6u4) between numerical variables. This particular step is crucial as it helps identify multicollinearity (when two features are highly correlated), which can distort model learning. Looking for correlations is specifically useful to see relationships like whether ‘Age’ is correlated with ‘BMI’ or ‘Price.’

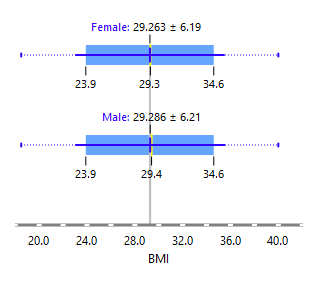
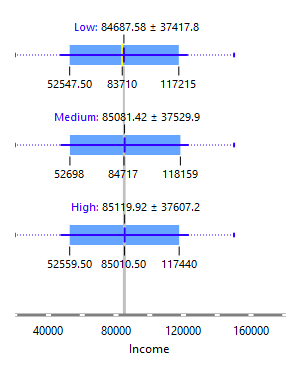
From [Figure 13](#bookmark=id.doc9q5n1n6u4), we can interpret that the number of prior claims increases, the predicted insurance Price increases. This is intuitive since more claims suggest higher risk. Higher deductible values may lead to lower insurance prices, as the insured is agreeing to pay more out-of-pocket, reducing risk for the insurer. Higher BMI is associated with increased insurance prices, possibly reflecting higher health risks. Older individuals may be priced higher due to increased risk. Slight relationship where wealthier customers may choose more comprehensive plans. However, Credit\_Score vs. Income and Credit\_Score vs. Deductible show mild correlations, and also Deductible vs. State\_Insurance\_Cost\_Index and Income vs. State\_Insurance\_Cost\_Index have small correlations (~+0.01), indicating minimal overlap. 

**Figure 13**

This step in our audit process is crucial to check fairness, bias, and transparency. as it ensures fairness by checking for strong correlations between sensitive attributes and price (like age and BMI), which essentially may warrant fairness audits. Highly correlated variables (e.g., Prior\_Claims and Price) may dominate model predictions, so it’s important to validate if they introduce bias. And correlation analysis gives auditors and stakeholders visibility into what influences pricing before diving into SHAP or LIME.

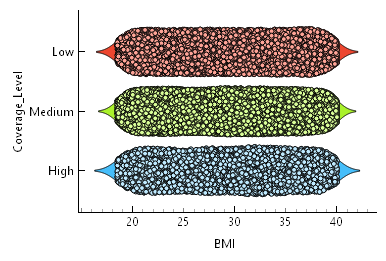
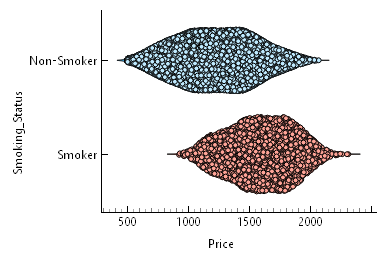
** **

**Figure 14 (boxplot of age groped by employment status)** **Figure 15 (boxplot of price grouped bu smoking status)**

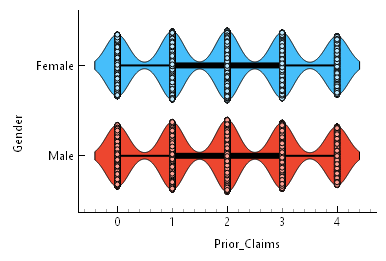
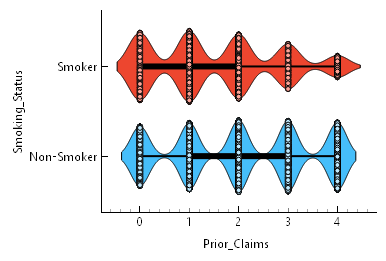
** **

**Figure 16 (Boxplot of BMI grouped by gender)** **Figure 17 (Boxplot of Income grouped by coverage level)**

The ‘Outliers’ widget automatically detects data points that deviate significantly from the rest. Identifying outliers in an audit is important because they can influence model decisions or distort fairness checks in the background. Then we use boxplots and violin plots to visualize the median, quartiles, and outliers of a feature grouped by a categorical variable. Boxplots are effective for understanding feature impact across categories; for example, how BMI differs by ‘Gender’ as in [Figure 16](#bookmark=id.lh3j44itlw81) or ‘Price’differs by ‘Smoking Status’ as in [Figure 17](#bookmark=id.lh3j44itlw81) Boxplots highlight outliers and spread, which could skew model learning. it is evident from [Figure 15](#bookmark=id.ty0h6kjw82ie), that boxplots showing that ‘Smokers’ consistently have higher Insurance ‘Prices’ than ‘Non-Smokers.’ A violin plot combines box plot and density plot, showing the distribution shape of a variable across categories. It gives a more detailed view than box plots about how feature values are spread. And also shows concentration and frequency of values in addition to medians and ranges. A violin plot snippet from our audit revealing that ‘non-smokers’ have a more tightly packed distribution of ‘Prices’ while ‘smokers’ have a wider spread, is shown in [Figure 19](#bookmark=id.ce90yyd8jlxv).

** **

**Figure 18 (violin plot of BMI vs Coverage level)** **Figure 19 (violin plot of price vs smoking status)**

** **

**Figure 20 (violin plot of prior claims vs gender)** **Figure 21 (violin plot of prior claims vs smoking status)**

This enables grouping or highlighting of potentially biased or sensitive attributes. This particular step helps auditor detect imbalances and understand correlations between input features and price See how attributes like age or smoking status influence pricing, and also identify outliers or extreme values. This helps auditors form hypotheses around potential bias or unfair pricing strategies.

**Apply the Same Preprocessing Steps**

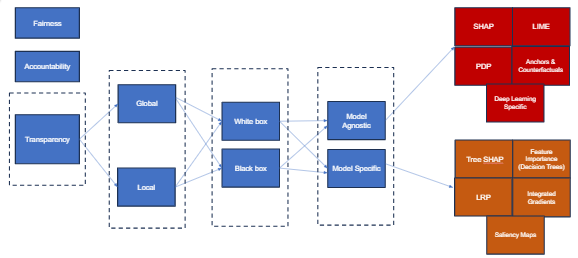
In our audit workflow ([Figure 24](#bookmark=id.v7qdiw2ydt6w)) , we made sure to replicate the exact same data preparation steps that were followed when the model was originally built by the insurer. This was important because feeding the model a differently preprocessed dataset could result in incorrect predictions or errors. So, just like the model’s training pipeline, we handled missing values using imputation, selected the appropriate features, scaled numerical data where needed, and ensured that all categorical variables were properly encoded. By mirroring the model’s data preprocessing pipeline, we maintained consistency and made sure the audit was fair, reliable, and based on the same assumptions the original model relied on.

**Load the .pkl Deployed Model**

For the "Model Retrieval and Target Preparation" step of our audit workflow ([figure 24](#bookmark=id.v7qdiw2ydt6w)), we began by loading the exact same .pkl model file that the insurer originally trained and deployed during Phase 1 ([figure 23](#bookmark=id.4bs1w342q7ft)) using the ‘Load model’ widget. Using the same saved model version was critical to ensure the audit remained valid and reflected how the model actually behaves in production. Once the model was loaded into Orange, we moved on to identifying and preparing the target variable. In this case, it is the insurance price. This involved clearly selecting "Prices" as the prediction output so that all evaluations, visualizations, and fairness assessments would be based on how accurately and consistently the model predicts that specific value. This setup laid the groundwork for a focused and meaningful audit.

**Check for Fairness, Transparency, and Accountability Using SHAP and Feature Importance**

The process we followed lines up well with the flow in the diagram ([Figure 22](#bookmark=id.c5z3sgua1ane)). We started by preparing the data using tools like the Discretizer to group continuous values like age and income into ranges, making it easier to compare across different people. Then, using the Domain Editor, we marked important features like smoking status or gender as sensitive, setting the stage for fairness checks. This aligns with the “Transparency” box in the image and helps us understand both the big picture (global view) and individual predictions (local view), shown in the next layer.

****

**Figure 22**

To dig into fairness and bias, we used the Fairness and Dataset Bias widgets. These helped us explore whether the model was treating all groups fairly or if some were being left out or favored, like seeing if smokers or people with certain health conditions were consistently charged more. These tools reflect the “Fairness” and “Accountability” goals at the start of the diagram ([Figure 22](#bookmark=id.c5z3sgua1ane)) .

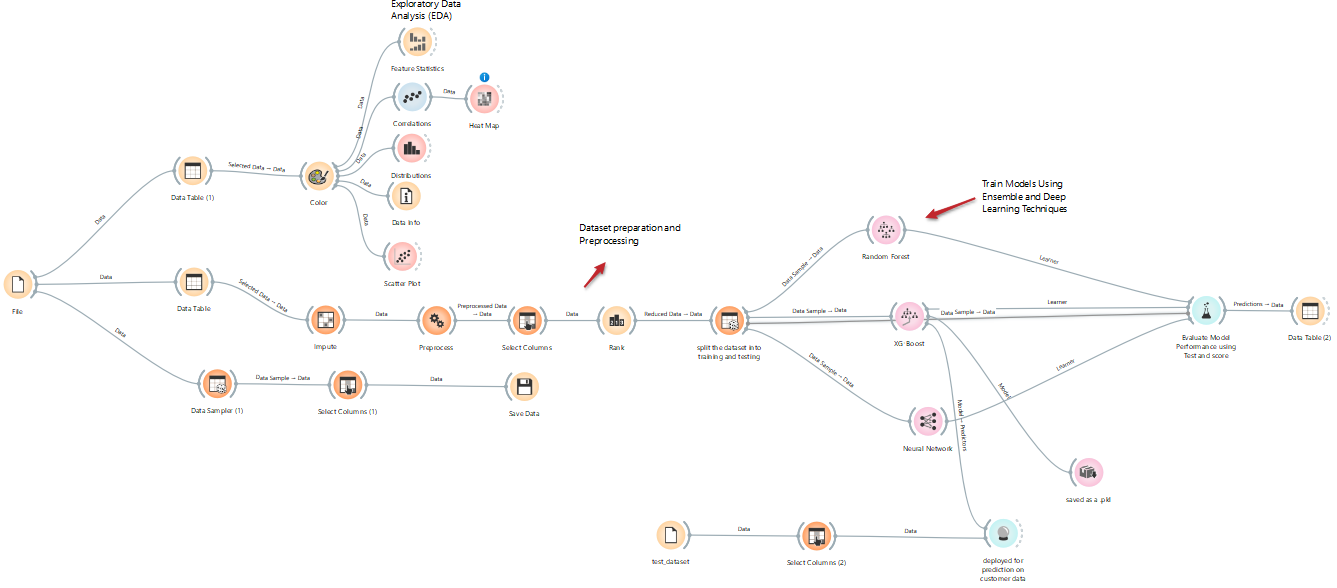
Next, to interpret how the model makes predictions, we used the Explain widget, which provides SHAP values and force plots. These tools support both global (overall model behavior) and local (individual prediction) explanations. Some of the techniques we used, like SHAP and permutation-based feature importance are model-agnostic, meaning they can explain any type of model. Others, such as tree SHAP, are model-specific and tailored to tree-based models like Random Forests or XGBoost. This flexibility allowed us to explain both black box models (which are complex and less transparent) and white box models (where internal workings are easier to understand). Altogether, this workflow helped us deeply understand, trust, and audit the model’s behavior, ensuring it aligns with ethical AI principles.

**Results**

As a result of building and auditing an insurance pricing model, the two Orange workflows offer a complete and approachable path from start to finish.

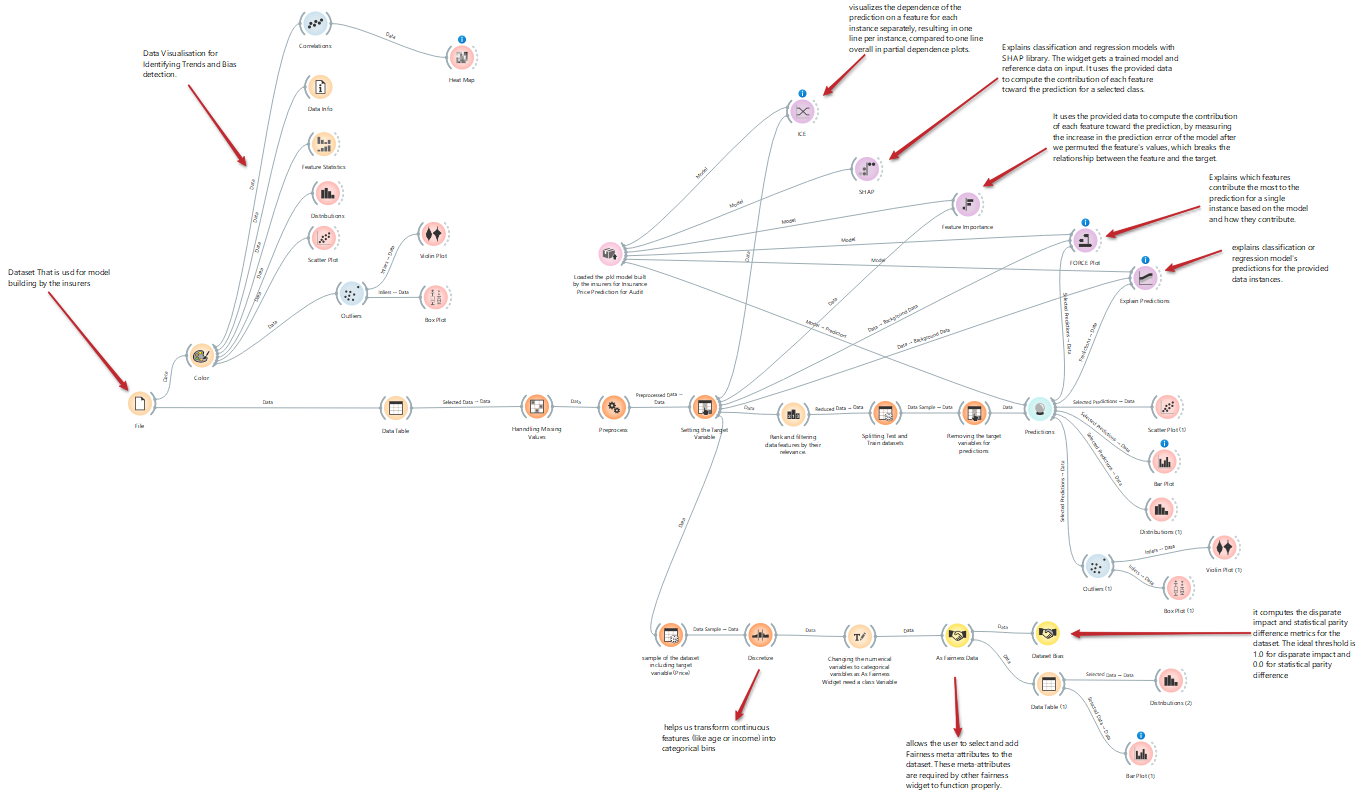
**Insurers Model Workflow in Orange**

The workflow shown in [Figure 23](#bookmark=id.4bs1w342q7ft) is a great example of how even people without deep programming skills can build powerful prediction models using Orange. Designed with a simple drag-and-drop interface, it allows users like insurance analysts or policy designers, to understand and experiment with data step by step. This visual pipeline breaks down complex tasks, such as selecting relevant information or comparing different algorithms, into clear building blocks. For insurers, it means they can explore how various customer features like age, income, or health conditions affect premium pricing, all without writing a single line of code. Orange also makes it easy to see which model performs best and helps with saving or deploying that model for real-world predictions. Ultimately, this workflow([figure 23](#bookmark=id.4bs1w342q7ft)) empowers domain experts to create and audit pricing systems confidently, using a tool that’s both transparent and user-friendly.

****

**Figure 23 (Insurers Model Workflow in Orange)**

**Audit Workflow in Orange**

****

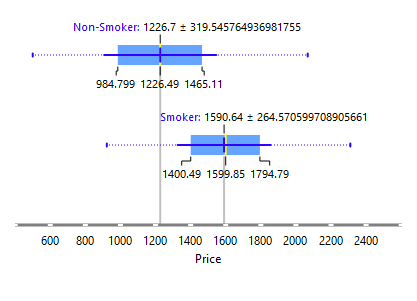
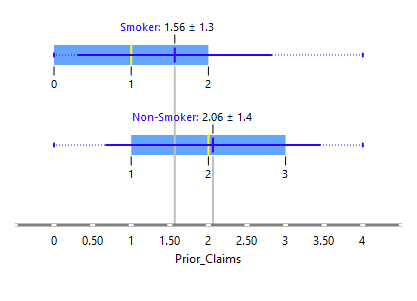
**Figure 24 (Audit Workflow in Orange)**

This workflow given in [Figure 24](#bookmark=id.v7qdiw2ydt6w) represents a mock audit of an insurance pricing model designed to evaluate fairness and explainability. It begins by loading a synthetic dataset and performing exploratory data analysis to detect patterns and potential biases. The pipeline includes preprocessing steps like handling missing values, defining the target variable, and discretizing continuous features. Explainability is assessed using SHAP, Force Plot, and Feature Importance widgets to understand how different features influence model predictions. The predictions are then examined using fairness widgets to simulate how the model treats different demographic groups. Overall, this mock audit simulates a responsible model evaluation process used by insurers.

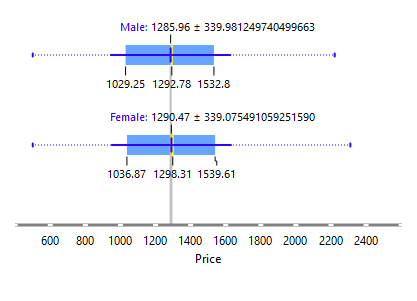
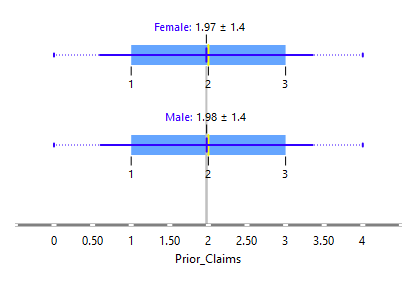
**Audit Report**

**Dataset Bias Analysis**

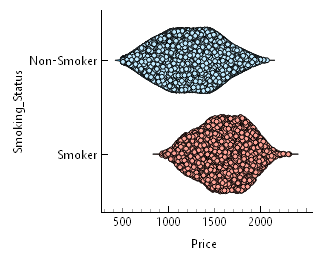
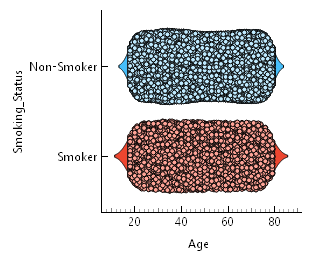
## **Bias Visualization, Group Comparisons:**

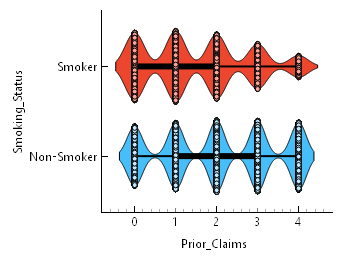
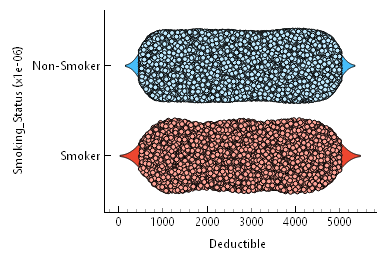
**Figure 25 (boxplot for price filtering smoking status)** **Figure 26 (boxplot for prior claims filtering smoking status)**

**Figure 27 (box plot for price filtering gender)** **Figure 28 (box plot for prior\_claims filtering gender)**

**Figure 29 (violin plot for price vs smoking status)** **Figure 30 (violin plot for age vs smoking status)**

**Figure 31 (violin plot for prior claims vs smoking status)** **Figure 32(violin plot for deductible vs smoking status)**

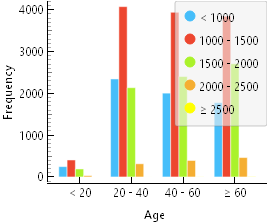
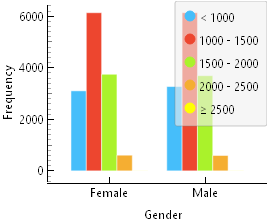
**Violin plotsin figures** [**29**](#bookmark=id.p83rq4pg27yj)**,** [**30,**](#bookmark=id.s0ltx8f4w7r1) [**31**](#bookmark=id.jqk9n3mtu4wa)**,** [**32**](#bookmark=id.5f6flp3xx8zw) **and boxplots in figures** [**25**](#bookmark=id.u2pb0dgddxo0)**,** [**26,**](#bookmark=id.dyirfxocf3a1)[**27**](#bookmark=id.zhs5iq3tpg0e)**, and** [**28**](#bookmark=id.lke6282rr4kr) for Smoking\_Status, Gender, and Pre\_Existing\_Conditions highlight:

1. Smokers consistently pay higher prices.
2. Minor gender price differences.
3. More prior claims and higher prices among smokers.
4. Deductibles tend to be slightly higher for smokers too.

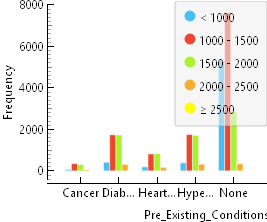
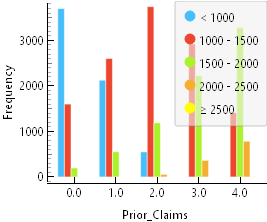
To ensure a fair and transparent model, we analyzed the distribution of the target variable (Price) across sensitive attributes like gender, smoking status, pre-existing conditions etc as shown below in figures [33](#bookmark=id.7mqwm62jvv7w), [34](#bookmark=id.dz0qw4ux3v4z),[35](#bookmark=id.wj94gn8o2j8z) and [36](#bookmark=id.11bfyjdg7aw2):

1. Gender Bias : No strong signs of pricing bias based on gender.
2. Smoking Status Bias : Smokers are consistently charged higher premiums than non-smokers. This aligns with risk-based pricing but presents a fairness trade-off; health risk vs affordability
3. Pre-Existing Conditions Bias : Customers with pre-existing conditions (Cancer, Diabetes, Heart Disease, Hypertension) tend to face **higher premiums** than those with “None”. Raises **ethical concerns** about pricing fairness for already vulnerable groups.
4. Employment Status Bias : Minor differences observed across employment types. Unemployed and retired individuals appear slightly disadvantaged.

**Fairness Audit: Disparate Impact and Statistical Parity**

**Figure 33 (bar chart of age vs price)** **Figure 34 (bar chart of gender vs price)**

**Figure 35 (bar chat of pre-existing conditions vs price)** **Figure 36 (bar chart of prior claims vs price)**

The fairness audit focused on various protected attributes mentioned in [table 3](#bookmark=id.e8h1g6c63x7v):

| Protected Attribute | Privileged Group | Disparate Impact | Statistical Parity Difference | Bias Summary |
| --- | --- | --- | --- | --- |
| Smoking\_Status | Non-Smoker | 0.628 / 1.594 | ±0.293 | Strong bias against smokers |
| Pre\_Existing\_Conditions | None | 0.718 – 1.275 | ±0.232 | Moderate bias |
| Age | 40–60 | ~1.000 | ~0.000 | No significant bias |
| Gender | Female | 0.99–1.01 | ±0.006 | |  | | --- |  | No significant bias | | --- | |
| Employment\_Status | Employed | ~1.000 | ~0.000 | No significant bias |

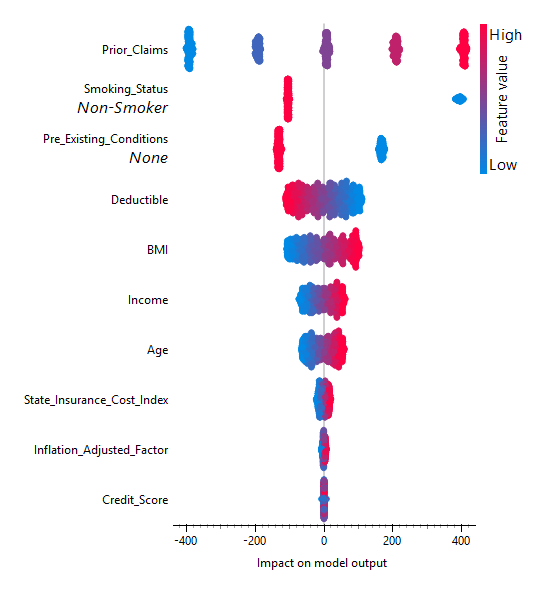
**Table 3**

1. Gender: The model performs equitably. Disparate Impact values are close to 1.0, and Statistical Parity Differences hover near 0, indicating no significant gender bias.
2. Age: Across different age groups, fairness metrics are within acceptable ranges. The model treats age groups consistently without favoring one over another.
3. Pre-Existing Conditions: This attribute shows notable bias. Individuals with conditions like diabetes, heart disease, or hypertension face higher prices, with Disparate Impact ranging from 1.2 to 1.27 and Statistical Parity Differences up to 0.16. This suggests unfair penalization of individuals with chronic health conditions, even if risk-based.
4. Smoking Status: Displays strong bias against smokers (Disparate Impact ~0.63, SPD ~0.29). While actuarially justified to some extent, the penalty may be disproportionate, especially when not balanced with other behaviors like exercise.

The model is fair with respect to gender and age, but biased against individuals with certain health statuses, particularly smokers and those with pre-existing conditions.

**Model Explainability and Feature Contributions (SHAP Analysis)**

1. **SHAP Summary Plot Analysis:**

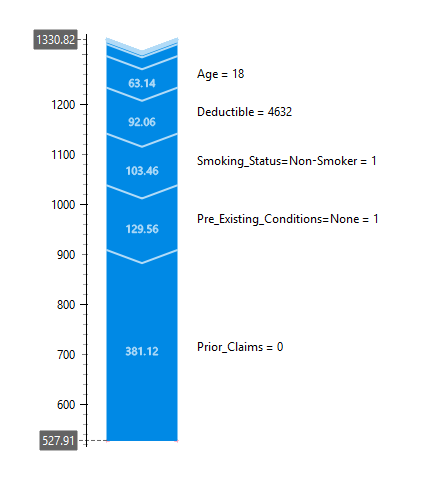


**Figure 37**

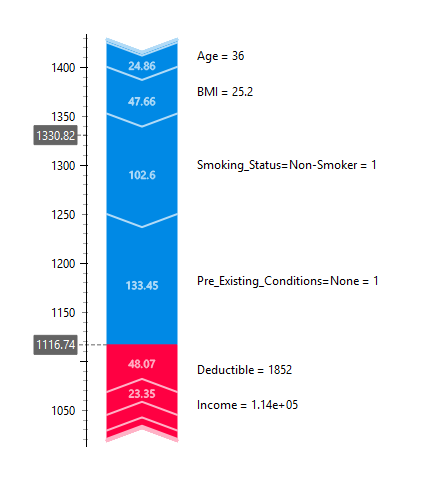
In this model, the features that stand out the most in influencing insurance pricing are Prior\_Claims, Smoking\_Status, and Pre\_Existing\_Conditions as noted in [figure 37](#bookmark=id.jvrxa8m2p1xf). The model tends to assign higher premiums to individuals with a history of more claims or those who are smokers. On the other hand, people who don’t smoke and have no pre-existing health conditions are typically predicted to have lower insurance costs. While other factors like age, deductible amount, BMI, and income also play a role in shaping the price, their impact is relatively smaller compared to the top three. This shows the model places significant weight on a person’s risk history and health-related behaviors when making predictions.

1. **Force Plot Insights:**

The two force plots below in [figure 38](#bookmark=id.ixb3t84p6x6j) and [figure 39](#bookmark=id.uimekxjh9f64) offer a clear and intuitive view of how different features affect individual insurance pricing predictions. In the first plot([figure 38](#bookmark=id.ixb3t84p6x6j)), a young individual (age 18) with no prior claims, a high deductible, and healthy indicators such as being a non-smoker and having no pre-existing conditions, sees their predicted price pushed significantly downward. All features contribute to lowering the cost, which aligns with expectations of lower risk. In the second plot ([figure 39](#bookmark=id.uimekxjh9f64)), while the person is also a non-smoker with no pre-existing conditions, their deductible is lower and income is relatively high, which slightly increases the prediction. Despite some cost-driving features, the healthy profile of the individual still helps keep the premium reasonable. These visualizations help us understand the rationale behind the model’s predictions and confirm that it responds to risk-related inputs as intended.



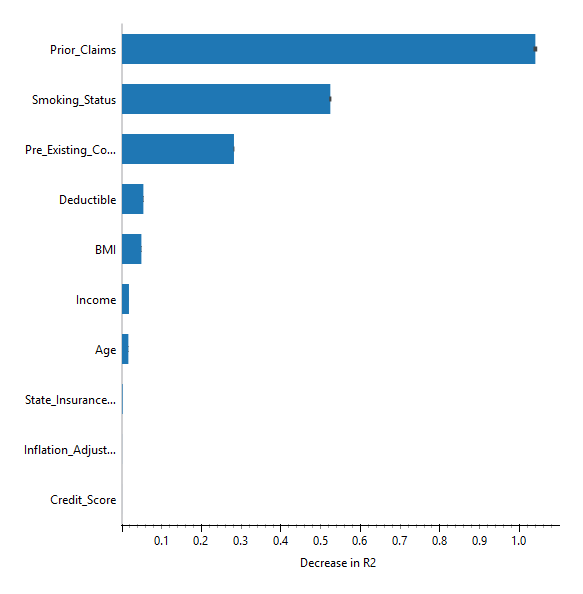
**Figure 38**



**Figure 39**

**Feature Importance (R² Drop Method)**

The feature importance analysis illustrated in [figure 40](#bookmark=id.f2ovj6jw2dne), derived from the permutation-based decrease in R² metric, provides critical insight into the factors most influential in determining insurance pricing predictions. Among all features, Prior\_Claims emerged as the most significant, with the greatest decrease in R² upon permutation, indicating that it is the most informative predictor of insurance cost. This is followed by Smoking\_Status and Pre\_Existing\_Conditions, which also exhibit substantial influence on model output. These results suggest that the model places considerable weight on historical risk indicators (e.g., prior claims) and health-related behavioral or status features (e.g., smoking, pre-existing conditions). In contrast, socio-economic variables such as Income and Credit\_Score, and demographic features like Age, have relatively lower importance in the model’s predictive logic. While this prioritization aligns with actuarial logic, it underscores the need to balance risk-based pricing with fairness considerations—especially given that features like smoking and health conditions may intersect with protected attributes. These findings further reinforce the necessity of bias auditing and the inclusion of fairness constraints during model development.



**Figure 40**

### **Key Ethical Principles Evaluation for XGBoost under Colorado Insurance Regulations**

#### **Fairness: Does the model treat people equitably regardless of sensitive characteristics?**

#### Fairness means the model should not disadvantage individuals based on protected characteristics like **age, gender, or health status**. According to Colorado’s insurance fairness rules, pricing models must avoid discriminatory outcomes, even when using actuarially relevant variables.

#### **Audit Insights on XGBoost**

#### **Gender:** The model passes fairness checks for gender. Disparate Impact scores remain close to 1.0, and Statistical Parity Differences hover near 0, suggesting no meaningful gender-based pricing bias.

#### **Age**: XGBoost treats different age groups with relative consistency. The fairness metrics remain within regulatory thresholds, and no single age bracket appears to be unfairly penalized.

#### **Pre-Existing Conditions**: Here, the model shows **significant fairness concerns**. Individuals with chronic illnesses like diabetes, hypertension, or heart disease face elevated prices (Disparate Impact 1.2–1.27, SPD up to 0.16), potentially crossing into regulatory non-compliance territory unless fully justified.

#### **Smoking Status**: Smokers face **notably higher premiums**, with Disparate Impact values as low as 0.63. While actuarially grounded, such penalties must be closely monitored to avoid disproportionately punishing behaviors that may be socioeconomic proxies or only partially related to risk.

#### **Transparency: Can decisions be explained clearly and understandably?**

#### 

#### Transparency is critical for regulatory compliance and public trust. It means stakeholders including regulators and customers, can understand how the model arrived at a decision.

#### 

#### **Audit Insights on XGBoost**

#### We used SHAP summary plots, force plots, and feature importance visualizations to break down both global and instance-level predictions.

#### SHAP summary plots showed that Prior\_Claims, Smoking\_Status, and Pre\_Existing\_Conditions were the top influencers of premium decisions.

#### Force plots made these explanations intuitive, showing how individual features pushed predictions higher or lower.

#### Permutation-based feature importance graphs provided a ranked view of how much each feature impacted model performance.

#### Together, these tools helped us demystify XGBoost, turning a black-box model into one that satisfies transparency expectations under Colorado law.

1. **Accountability: Can the model’s decisions be traced, justified, and audited?**

Accountability means that every prediction made by the model can be traced back to clear logic and data. It also ensures that decisions can be justified in legal or regulatory reviews.

**Audit Insights on XGBoost**

* 1. Through tools like the Fairness widget and Dataset Bias explorer, we traced how predictions varied by subgroup.
  2. Visual tools such as box plots, violin plots, and SHAP-based breakdowns, allowed us to link pricing decisions directly to input features.
  3. This traceability supports the level of justification required by Colorado regulations, making it easier for insurers to respond to external audits or customer concerns.

### **Final Recommendations**

Our audit revealed that the data used to train the insurance pricing model contains subtle yet meaningful biases. Especially around age groups, smoking status, and health conditions. These biases can lead to unfair pricing for certain groups. To address this, we recommend using techniques like reweighting or sampling adjustments to balance the representation of all groups. By reducing imbalances in the training data, we can help ensure that no group is unintentionally favored or penalized.

#### **1. Actively Address Data Bias**

The training data used in the model shows imbalances across age, health status, and smoking groups. These biases can cause the model to systematically overprice certain individuals. We recommend using bias mitigation techniques like reweighing, sampling adjustment, or fair preprocessing to ensure group parity is not violated.

#### **2. Integrate Fairness Checks Throughout the Lifecycle**

Fairness must not be an afterthought. Using tools like Orange’s Fairness widget, audits should be baked into the full model pipeline from data cleaning to deployment. This ensures the model continuously complies with anti-discrimination laws and internal ethical standards.

#### **3. Use SHAP and Force Plots to Explain Decisions**

To meet transparency mandates, make SHAP and force plots a standard part of model review. These tools enable both technical and non-technical audiences to understand and challenge predictions, which aligns with the regulation’s call for interpretability.

#### **4. Document Every Decision for Strong Accountability**

A compliant model requires a strong audit trail. By maintaining records of all fairness checks, feature importance decisions, and model revisions, insurers can justify any outcome and demonstrate good-faith efforts to comply with fairness laws.

#### **5. Reevaluate Use of Sensitive Predictors**

While features like **smoking status** and **pre-existing conditions** are statistically relevant, relying heavily on them can create ethical and regulatory challenges. We suggest reevaluating their weight in the model, especially when they result in disproportionate pricing penalties. Contextualizing these inputs, perhaps with behavioral or preventive health data, could support fairer outcomes.

**References**

1. [**https://www.hhs.gov/guidance/document/national-health-expenditure-nhe-fact-sheet**](https://www.hhs.gov/guidance/document/national-health-expenditure-nhe-fact-sheet)

1. [**https://www.datasciencecentral.com/how-to-explore-historical-data-patterns-with-machine-learning/**](https://www.datasciencecentral.com/how-to-explore-historical-data-patterns-with-machine-learning/)

1. [**https://www.explorationpub.com/Journals/edht/Article/10119**](https://www.explorationpub.com/Journals/edht/Article/10119)

1. [**https://securiti.ai/nyc-automated-employment-decision-tool-aedt/**](https://securiti.ai/nyc-automated-employment-decision-tool-aedt/)

1. [**https://www2.deloitte.com/us/en/pages/risk/articles/ai-regulation-colorado-dept-of-insurance.html**](https://www2.deloitte.com/us/en/pages/risk/articles/ai-regulation-colorado-dept-of-insurance.html)

1. [**https://orangedatamining.com/widget-catalog/data**](https://orangedatamining.com/widget-catalog/data)

1. [**https://gretel.ai/**](https://gretel.ai/)

1. [**https://shap.readthedocs.io/en/latest/index.html**](https://shap.readthedocs.io/en/latest/index.html)

1. [**https://shap.readthedocs.io/en/latest/example\_notebooks/api\_examples/plots/bar.html**](https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/bar.html)