

Premium Pricing Predictor (PPP)

1.0 Abstract

The next step in socially equitable life insurance is life insurance policies independent of health preconditions. Our novel intelligent solutions determine life insurance premiums with powerful accuracy without user health preconditions, demonstrating 99% accuracy in premium calculation across all four of Vitech's Plans: Platinum, Gold, Silver, and Bronze. Our systems were trained with a robust dataset of over 500,000 cases and tested on over 100,000 cases. Our analytics provide next-generation tools with the capacity to disrupt life insurance: a clear example of artificial intelligence as an asset, not a danger, to society.

1.1 Novelty of Approach and Relevance

Whereas medical preconditions are evidently well documented in Vitech's Database and API which include risk specifications for the following factors:

- Chronic viral hepatitis B without delta-agent
- Ataxic cerebral palsy
- Diarrhea, unspecified
- Tachycardia, unspecified
- Obstructive Sleep Apnea
- Unspecified fracture of specified metacarpal bone with unspecified laterality
- Other abnormalities of heart beat
- Type 2 diabetes mellitus with hyperglycemia
- HIV disease resulting in other bacterial infections
- Cough with hemorrhage

Medical preconditions have been the target of public administrations and policy aimed to enable insurance for a larger proportion of the population. Given the historical precedence of the Affordable Care Act, preventing Health Care insurers from charging premiums based on pre-conditions, Life Insurance Companies remain under threat by potential policy measures.

Furthermore, providing insurance premiums without accounting for pre-conditions helps enhance the image of a life insurance company, simultaneously encouraging more individuals to purchase premiums from the company. Advantages include:

- Reduction in Time and Complexity to Calculate Premiums
- Decreased Risk to Public Policy Changes
- Better Image increasing Potential Clients

A pivot towards premium and plan prediction without the use of client medical precondition is therefore necessary based on future expectations and beneficial in the short term.

2.0 Methods and Interface

Data Cleansing	<ul style="list-style-type: none">• Vitech API• R / SQL• Microsoft Azure
Data Visualization	<ul style="list-style-type: none">• Tableau• Python
Machine Learning	<ul style="list-style-type: none">• Python• Microsoft Azure
Front End	<ul style="list-style-type: none">• HTML / CSS• Bootstrap• Django

2.1 Data Cleansing

To maintain the nature of the data, a plurality of queries based on the Vitech API, were used to comprehensively access a balanced data set in terms of individuals from different states, with difference combinations of smoking / number of dependents / marital status, and gender. The balanced nature of the data set consisted of over 600,000 rows of data.

2.2 Data Visualization

Tableau was utilized to create concise, yet impactful visualizations of the overall data set and variable trends relevant to understanding the impact of sex, tobacco use, geography, and age with the selection of premium pricing. Alternatively, Python based packages were used to visualize errors and dimensionality reduction from machine learning algorithms.

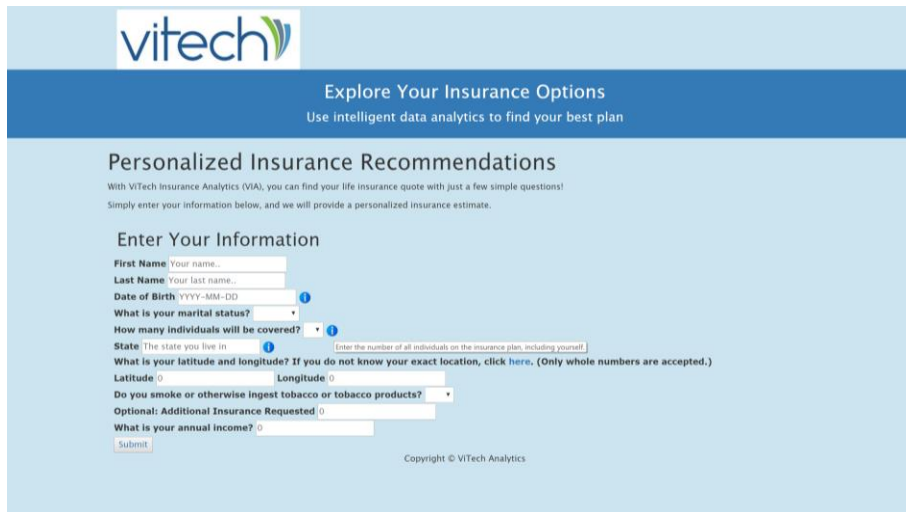
2.3 Machine Learning

Microsoft Azure's Machine Learning Studio was used to run basic artificial intelligence experiments involving multiclass neural networks and boosted decision tree forests for classification of premium purchase and regression of premium pricing respectively. Based on results of Azure experiments, subsequent studies were pursued for confirmation and optimization using Python's Scikit-Learn functionality.

We largely utilized Scikit-Learn for precise premium pricing regression, training models to very accurately predict premium pricing for each plan based on personal information. Additionally, we made use of Scikit-Learn's support vector machine (SVM) and decision tree forest functionality for classification purposes, matching clients to their most suitable plans.

2.4 Front End

Our frontend code utilizes HTML and CSS to format the form that will determine the predicted plan recommendation. The first page contains a form that includes various kinds of input, such as text fields and dropdown menus. The input fields all have placeholders or blank areas to signal to the user to replace the empty content. There are blue information icons that the user can hover over to find more information on how to answer the questions. The page also includes a hyperlink that opens in a new tab to assist the user in determining the latitude and longitude of their location.



The screenshot shows the 'Explore Your Insurance Options' page. It features a header with the Vitech logo and a sub-header 'Explore Your Insurance Options' with the tagline 'Use intelligent data analytics to find your best plan'. The main section is titled 'Personalized Insurance Recommendations' and includes a brief explanation: 'With ViTech Insurance Analytics (VIA), you can find your life insurance quote with just a few simple questions! Simply enter your information below, and we will provide a personalized insurance estimate.' Below this is a form titled 'Enter Your Information' with the following fields: 'First Name' (placeholder: Your name...), 'Last Name' (placeholder: Your last name...), 'Date of Birth' (placeholder: YYYY-MM-DD), 'What is your marital status?' (dropdown menu), 'How many individuals will be covered?' (dropdown menu), 'State' (placeholder: The state you live in), 'What is your latitude and longitude? If you do not know your exact location, click here. (Only whole numbers are accepted.)' (with a link), 'Latitude' (placeholder: 0), 'Longitude' (placeholder: 0), 'Do you smoke or otherwise ingest tobacco or tobacco products?' (dropdown menu), 'Optional: Additional Insurance Requested' (placeholder: 0), and 'What is your annual income?' (placeholder: 0). A 'Submit' button is at the bottom left, and 'Copyright © ViTech Analytics' is at the bottom right.

After the "Submit" button is pressed, the user is then redirected to the plan comparison page, which also includes the recommended insurance plan. The header section includes a Bootstrap navbar that allows the user to quickly skip to the section of the page that they are most interested in.



The screenshot shows the 'Your Insurance Options' page. It features a header with the Vitech logo and a sub-header 'Your Insurance Options' with the tagline 'Compare coverage and prices side by side'. The main section is titled 'Platinum, Gold, Silver and Bronze Plans Information'. The page also includes a Bootstrap navbar with 'Options' and 'Recommendations' links.

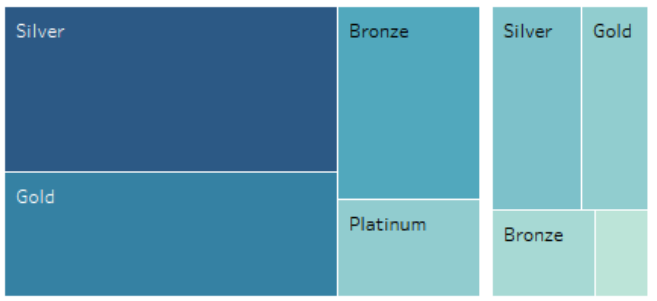
The second page also includes a table that includes the information for each plan.

Your Insurance Options				
Compare coverage and prices side by side				
Platinum, Gold, Silver and Bronze Plans Information				
Plan Type	Base Price	AD&D ⓘ	Deductible	Coverage Amount
	\$110	\$50,000	\$3,000	\$100,000
	\$70	\$30,000	\$5,000	\$100,000
	\$40	\$20,000	\$7,000	\$100,000
	\$20	\$10,000	\$8,000	\$100,000
Our Insurance Recommendations				

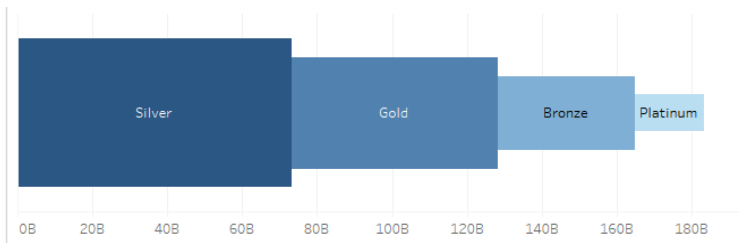
Last but not least, the logo at the top right on both pages can be clicked if the user is interested in finding out more about ViTech!

3.0 Technical Findings

3.1 Visualizations and Demographic Insights



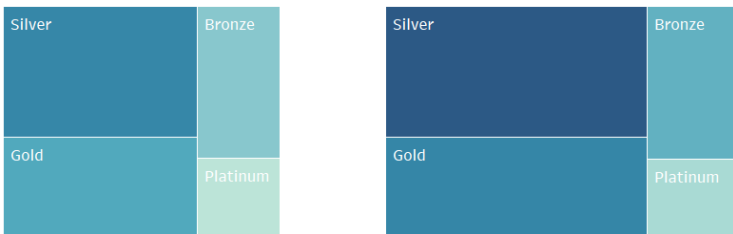
This graphics reflects the comparative distributions between non tobacco users (left) and tobacco users (right) in terms of purchasing of plans and their relative quantities.



The graphic depicts the effect of summation of total additional insurance by purchasing plan, where purchasers of the Silver plan were responsible for the largest purchase of additional insurance.



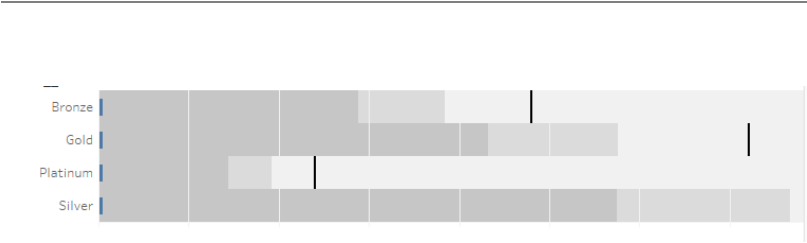
The graphic depicts the overall relative distributions of the four main premium plans offered, indicating the relative prominence of gold and silver plans among clients.



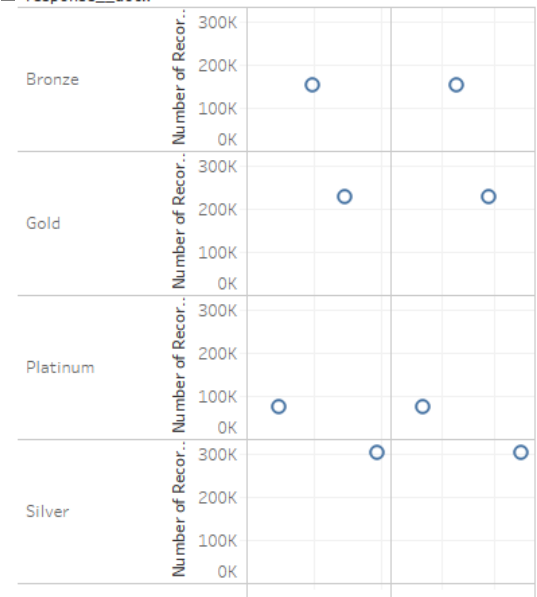
The graphic depicts the comparative distributions of plans purchased between married (left), and single (right). Note that Silver and Gold account for a greater proportion of premium purchases among those who are single, both the relative differences are very small.

response___	response__docs___	No	Yes
M	Bronze	<div></div>	<div></div>
	Gold	<div></div>	<div></div>
	Platinum	<div></div>	<div></div>
	Silver	<div></div>	<div></div>
S	Bronze	<div></div>	<div></div>
	Gold	<div></div>	<div></div>
	Platinum	<div></div>	<div></div>
	Silver	<div></div>	<div></div>

The graphic depicts the comparative relationships between the highly multidimensional interactions of tobacco use, marital status, and plan selection by quantification. Evidently single non-smokers are more prominent, whereas smokers in marriage are more common amidst clients.



The graphic depicts the comparative quantities of individuals purchasing of each of the four key plan types, whereas silver accounts for the greatest sum of annual income in terms of number of customers with relative affluence.



The graphic depicts the tendency for height to weight ratios to vary greatly between single and married individuals between the four plans. Whereas single and married individuals in the same plan have highly similar height and weight ratios, it is observed that the weight and height in terms of lack of proportionality and unhealthiness is highest among the gold and platinum subgroups.

3.2 Microsoft Azure Experiments

Microsoft Azure's Machine Learning Studio was utilized to test the capacity of a Boosted Regression Tree trained on data excluding preconditions to predict premium pricing for each of the four key premium types. Application of the tree found a mean squared error below 2 dollars for the pricing prediction of each of the four premium plans.

Hyperparameters of Boosted Regression Tree

- Maximum number of leaves per tree: 20
- Minimum number of samples per leaf node: 10
- Learning rate: 0.2
- Total number of trees constructed: 100

3.2.1 Azure Experiments with Health Preconditions





3.2.2 Azure Experiments with Health Preconditions

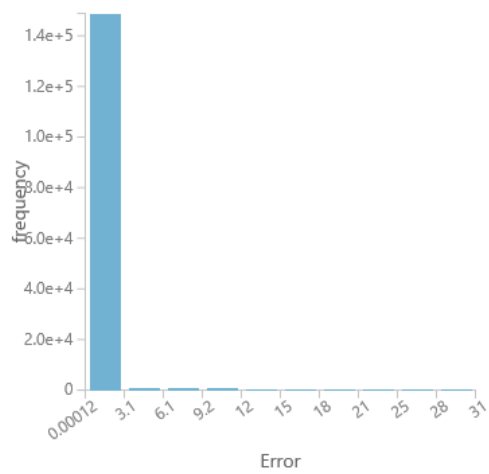
Azure Experiments training the boosted random tree decision forest without health preconditions demonstrates minimal differences in terms of the mean absolute error amounting to less than 10 cents across all four of the plans. These findings are highlighted through a comparative analysis of the provided data.

Platinum

Metrics

Mean Absolute Error	0.559877
Root Mean Squared Error	1.756054
Relative Absolute Error	0.008349
Relative Squared Error	0.00053
Coefficient of Determination	0.99947

Error Histogram

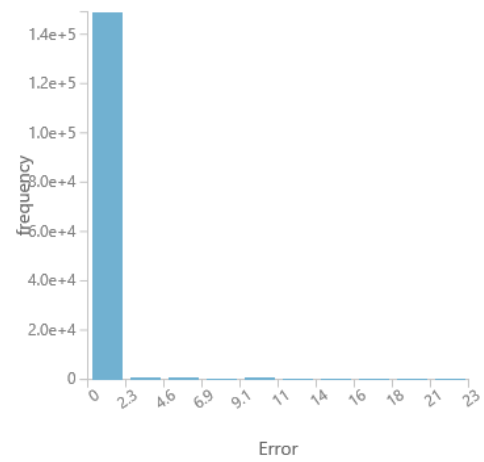


Gold

Metrics

Mean Absolute Error	0.471699
Root Mean Squared Error	1.253212
Relative Absolute Error	0.010882
Relative Squared Error	0.000633
Coefficient of Determination	0.999367

Error Histogram

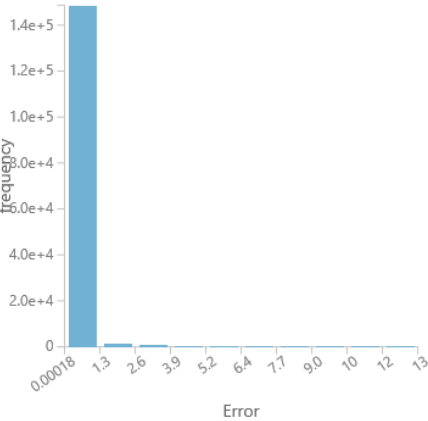


Silver

Metrics

Mean Absolute Error	0.329483
Root Mean Squared Error	0.86687
Relative Absolute Error	0.012613
Relative Squared Error	0.000806
Coefficient of Determination	0.999194

Error Histogram

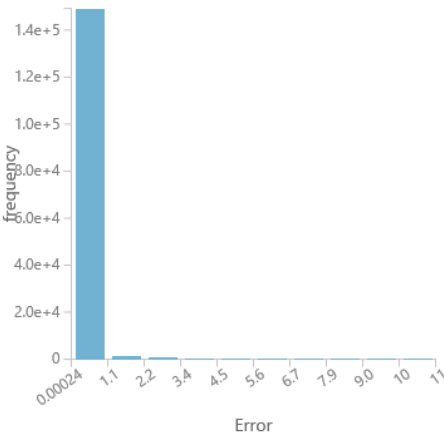


Bronze

Metrics

Mean Absolute Error	0.256168
Root Mean Squared Error	0.671514
Relative Absolute Error	0.016449
Relative Squared Error	0.001266
Coefficient of Determination	0.998734

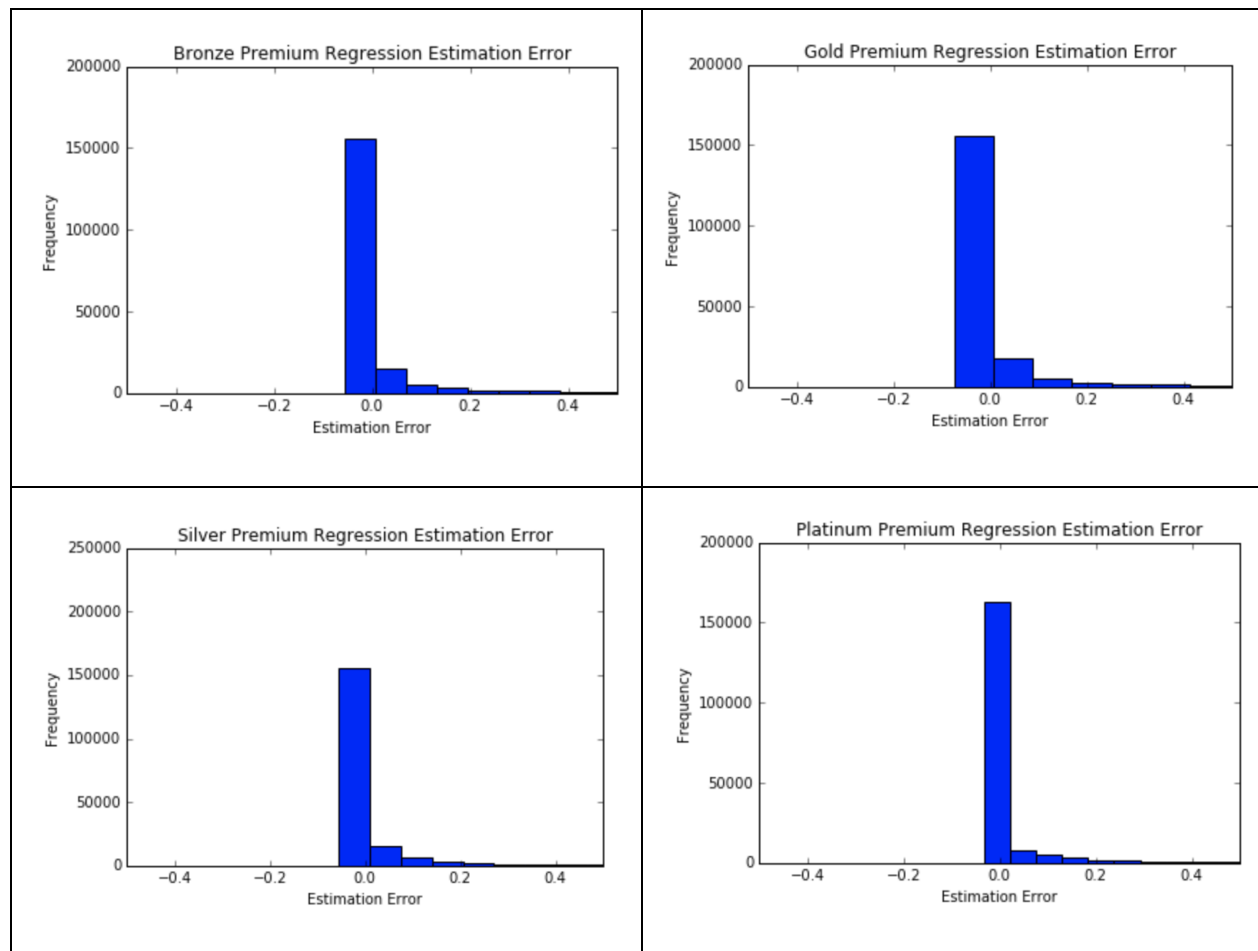
Error Histogram



3.3 Python-Based Regression Algorithms

Plan	Mean-Squared Error
Bronze	0.3087
Silver	0.3190
Gold	0.3213
Platinum	0.3187

Regression Error Histogram



Sample: Machine Learning for Prediction of Gold Premium Prices

```
# Import required packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Display plots within the notebook
%matplotlib inline
```

```
data_tr = pd.read_csv("data/FullDatasetVitech.csv", dtype=str, sep=',', index_col=0)
data_tr.head()
```

```
x_tr = data_tr.values[:, 4:].astype(float)
y_tr = data_tr.values[:, 2].astype(float)
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x_tr, y_tr, test_size=0.25, random_state=0)
```

```
clf = RandomForestRegressor(n_estimators=500, random_state=0)
clf.fit(X_train, y_train)
ranking = np.argsort( clf.feature_importances_ )[::-1]
```

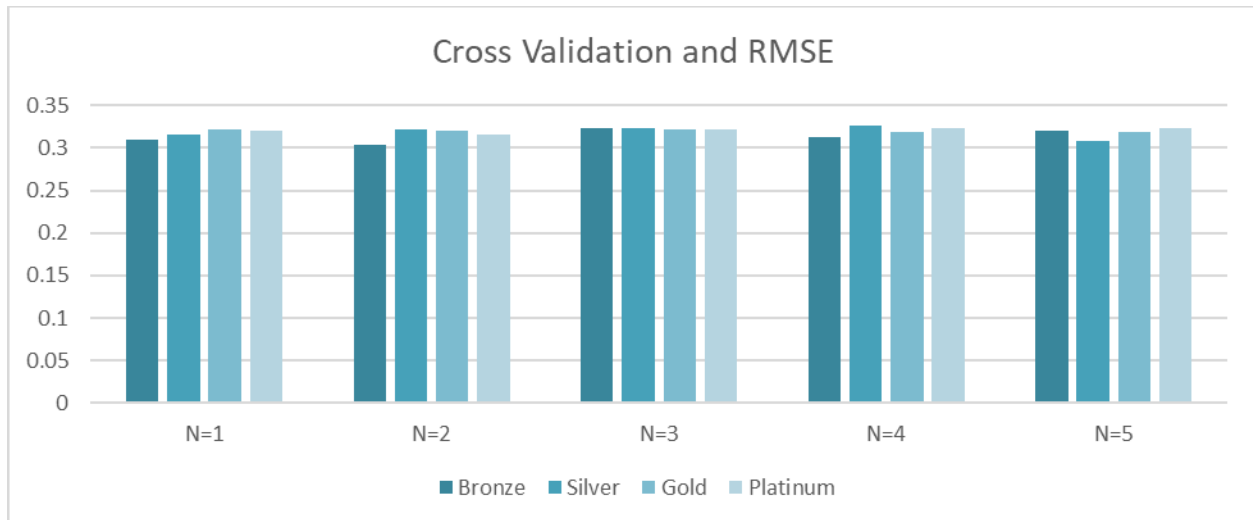
```
y_pred = clf.predict(X_test)
from sklearn.metrics import mean_squared_error
mean_squared_error(y_test, y_pred, sample_weight=None, multioutput= 'uniform average
')
```

3.4 Cross Validation

Use of stratified cross validation across 5 folds with our data pipeline, confirmed the accuracy of our root mean square error metrics derived with a sample run. There were consistent results

Plan	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$
<i>Bronze</i>	0.310	0.304	0.323	0.313	0.320
<i>Silver</i>	0.316	0.322	0.324	0.327	0.308
<i>Gold</i>	0.322	0.320	0.322	0.319	0.318
<i>Platinum</i>	0.320	0.316	0.321	0.323	0.324

between samples among all four runs. Cross validation experiments, therefore prove the accuracy of our modelling techniques.



3.5 Classification Algorithms

We tackled the task of recommending optimal plans to clients by, based on personal details, classifying each person into the bronze, silver, gold, or platinum category. Choosing an insurance plan is, unfortunately, subject to more human bias and subjectivity than determining premium pricing, but our models nevertheless made consistently reliable predictions. We found that, using both support vector machines and random decision tree forests for classification, we were able to achieve reasonable prediction accuracy upwards of 90%. Plan classifications disagreed strongly with purchased plans (e.g. predicted: bronze, purchased: platinum) for only 0.12% of clients.

4.0 Summary

Criteria	Our Approach
Novelty	<ul style="list-style-type: none">• Focusing on accurate premium prediction without health preconditions is an unexplored, yet relevant topic• Providing analytics for company purposes while simultaneously developing UI for clients• Civic Impacts acting as the analytic basis of an Affordable Life Insurance movement
Technical Difficulty	<ul style="list-style-type: none">• Use of versatile machine learning models including boosted decision tree forests, regression, and support vector machines• Multiple platforms for machine learning including Microsoft Azure and Python• Use of cross validation and hyper parameterization to optimize machine learning algorithms• Visual analytics using professional grade software such as Tableau to draw graphical insights
Usability	<ul style="list-style-type: none">• Includes a user-friendly, visual interface for entering personal information, reviewing insurance options, and receiving personalized recommendations• Minimization of form entry components for fast quote analysis• Integrates front-end with functionality to add intelligent back-end for ease of use
Insightful Findings	<ul style="list-style-type: none">• High level of accuracy can be maintained in quote development without accessing user pre-conditions• Impacts of finding include increased customer satisfaction and appeal, decrease in quote complexity, and decreased susceptibility to public policy