Project Title:

Comparative Analysis of Classification Algorithms for Predicting Customer Insurance Purchases

Your Name:

M.vani sree

Date:

29 June 2025

Abstract

In this project, we explore a real-world case study of predicting customer insurance purchases based on demographic features such as age and estimated salary. As an analyst at a bank insurance company, the objective was to apply multiple classification algorithms to accurately forecast customer behavior. A dataset derived from social network ads was used to train and evaluate various machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, and Random Forest classifiers.

Each model was implemented separately and compared based on performance metrics such as accuracy, confusion matrix. The results were visualized using decision boundaries to understand the impact of age and salary on insurance purchase decisions. Predictions for specific age and salary inputs were also conducted. The study concludes with the identification of the best-performing algorithm that balances accuracy with generalization, ensuring minimal overfitting. This project provides valuable insights into model selection and application of AI in the insurance sector.

Introduction

In today's data-driven economy, understanding customer behavior is essential for organizations seeking to remain competitive. The insurance industry, in particular, relies heavily on data analytics to identify potential clients and tailor offerings to meet customer needs. One of the core challenges in this space is predicting whether a prospective customer is likely to purchase an insurance policy based on demographic and financial indicators.

This project aims to address this challenge by leveraging artificial intelligence and machine learning techniques to predict customer insurance purchases using historical data. Specifically, the dataset includes two key features — Age and Estimated Salary — and a binary target indicating whether or not the customer purchased insurance.

The primary goal is to build and compare various classification models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, and Random Forest. These models will be trained, evaluated, and compared based on metrics such as accuracy, confusion matrix.

By identifying the most effective model, the project seeks to support better decision-making in customer targeting and policy recommendations. This AI-based approach not only improves the accuracy of predictions but also enhances the bank insurance company's ability to understand the underlying patterns in customer behavior.

Literature Review

Predictive analytics has become a cornerstone in the insurance industry, where understanding customer behavior can significantly influence marketing strategies and risk assessment. Several research studies have explored the use of machine learning techniques to predict customer decisions and insurance purchases.

A study by **Chaudhuri et al. (2019)** emphasized the effectiveness of logistic regression in binary classification problems involving demographic data. Their results showed that while logistic regression offers interpretability, its performance may be limited in capturing non-linear relationships.

Kumar and Ravi (2016) conducted a comprehensive survey on classification models used for customer churn prediction in financial services. They highlighted the growing importance of ensemble techniques such as Random Forests and their ability to reduce overfitting and improve accuracy in real-world datasets.

Zhang et al. (2018) demonstrated that support vector machines can outperform traditional models when the dataset has well-separated classes. Their work also emphasized the role of proper scaling and kernel selection in SVM models.

Recent developments in **AI for financial services** have shown that combining multiple models and comparing them using appropriate performance metrics is an effective strategy for choosing optimal solutions. This comparative approach allows organizations to balance complexity, training time, and prediction accuracy.

By reviewing these studies, it becomes clear that no single algorithm is universally superior. Instead, performance varies depending on the data distribution, feature selection, and the presence of noise. This understanding forms the foundation of our comparative study on insurance purchase prediction.

Problem Statement

In the highly competitive insurance industry, accurately identifying potential customers who are likely to purchase insurance policies is a critical business objective. The primary challenge lies in building a reliable classification system that can generalize well to new, unseen data without overfitting.

This project aims to tackle this challenge by analyzing a dataset comprising user details, specifically **age** and **estimated salary**, to predict insurance purchase decisions. The objective is to apply multiple supervised machine learning classification algorithms and conduct a thorough comparative analysis based on standard evaluation metrics such as **accuracy**.

A core focus of this study is to identify an optimal model that provides high predictive performance while maintaining simplicity and interpretability. The project explores the trade-offs between different algorithms in terms of model complexity, training time, and overfitting risk. Ultimately, the goal is to select a model that balances precision and generalization, thereby enabling informed, data-driven decision-making for targeted marketing and insurance offerings.

Data Collection and Preprocessing

The dataset used for this project is titled **Social_Network_Ads.csv**, a common dataset used to model consumer behavior in digital marketing contexts. It contains user data such as age, estimated salary, gender, and purchase status. However, for this study, only **Age** and **Estimated Salary** were selected as features, as they are directly relevant to the objective of predicting insurance purchase decisions.

The dataset comprises the following:

- **Features (X):**
    - Age
    - Estimated Salary
- **Label (Y):**
    - Purchased (0 = No, 1 = Yes)

Before model training, the dataset underwent the following preprocessing steps:

1. **Feature Selection:** Only the age and estimated salary columns were selected for input.
2. **Data Splitting:** The dataset was split into **training** and **test sets** using a 75-25 split via train_test_split from sklearn.model_selection.
3. **Feature Scaling:** Standardization was applied to both the training and test sets using StandardScaler to ensure all features had a mean of 0 and a standard deviation of 1. This step was especially important for algorithms such as SVM and KNN, which are sensitive to feature magnitudes.

4. **Target Encoding:** As the target column was already binary (0 or 1), no additional encoding was required.

The preprocessing ensured that the data was normalized, balanced, and suitable for training a variety of machine learning classification models.

Methodology

To predict whether a customer will purchase insurance based on their age and estimated salary, a comparative analysis was conducted using five supervised machine learning classification algorithms. Each model was trained on the same preprocessed dataset and evaluated using consistent performance metrics. The following algorithms were selected:

1. Logistic Regression

A statistical model used for binary classification. It outputs a probability between 0 and 1, ideal for predicting binary outcomes such as purchase decisions. It serves as a strong baseline model due to its simplicity and interpretability.

2. K-Nearest Neighbors (KNN)

A non-parametric algorithm that classifies data points based on the labels of their closest neighbors. The value of **K** (number of neighbors) was tuned during experimentation. KNN is useful for problems where decision boundaries are nonlinear and data points exhibit localized patterns.

3. Support Vector Machine (SVM)

A powerful classification algorithm that finds the optimal hyperplane that best separates the two classes. SVM can handle both linear and non-linear decision boundaries using kernel tricks. Feature scaling was essential to ensure optimal performance of SVM.

4. Decision Tree Classifier

This algorithm builds a flowchart-like tree structure to make decisions. It splits data based on feature thresholds to classify outcomes. Decision Trees are intuitive, easy to visualize, and work well for small-to-medium datasets.

5. Random Forest Classifier

An ensemble technique that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Random Forest offers higher robustness and generalization capability compared to individual decision trees.

Each model was trained on a standardized dataset using the same training-test split. During evaluation, metrics such as **accuracy**, **confusion matrix**. were computed. Visualization of decision boundaries was also performed to interpret how each model distinguishes between buyers and non-buyers based on age and salary.

The comparative nature of this methodology allows us to evaluate which model is best suited for this prediction task by analyzing performance, simplicity, overfitting risk, and computational efficiency.

Implementation

The project was implemented using the **Python programming language** in the Jupyter Notebook environment. The following libraries were used for data manipulation, model development, and visualization:

- pandas and numpy for data handling
- matplotlib and seaborn for visualization
- sklearn for machine learning models and evaluation metrics

Each classification algorithm was implemented in a **separate Python file** to ensure modularity and better debugging. Below is a summary of the implementation steps for each model:

1. Logistic Regression

- Standardized the features using StandardScaler
- Trained LogisticRegression() from sklearn.linear_model
- Plotted decision boundaries
- Predicted insurance purchase for test cases

2. K-Nearest Neighbors (KNN)

- Standardized inputs using StandardScaler
- Trained model using KNeighborsClassifier(n_neighbors=5)
- Plotted decision boundary and confusion matrix
- Tested model on multiple new inputs

3. Support Vector Machine (SVM)

- Used SVC(kernel='linear') with scaled features
- Created contour plots to visualize class separation
- Performed predictions on custom age-salary pairs

4. Decision Tree Classifier

- Implemented using DecisionTreeClassifier(criterion='entropy')
- Did not require feature scaling
- Used tree-based splits to classify purchase decisions

5. Random Forest Classifier

- Used RandomForestClassifier(n_estimators=10, criterion='entropy')
- Built an ensemble of decision trees
- Visualized predictions and accuracy results

GitHub Link of the project :

https://github.com/mVanisree/Customer-Insurance-Purchase-Prediction

Project *: ..\ARTIFICIAL INTELLIGENCE\MODELS\Insurance-Prediction-Project\Logistic Regression\Logistic Regression .ipynb*

..\ARTIFICIAL INTELLIGENCE\MODELS\Insurance-Prediction-Project\KNN\KNN MODEL.ipynb

..\ARTIFICIAL INTELLIGENCE\MODELS\Insurance-Prediction-Project\SVM\SVM.ipynb

..\ARTIFICIAL INTELLIGENCE\MODELS\Insurance-Prediction-Project\Desicion Trees\decision trees.ipynb

..\ARTIFICIAL INTELLIGENCE\MODELS\Insurance-Prediction-Project\RandomForest\RandomForestClassifier.ipynb

Results

Each model was trained and evaluated on the same preprocessed dataset split (75% training, 25% testing). Below is a comparison of their performance:

Comparative Accuracy Table

Model	Accuracy	Inferences
Logistic Regression	89%	Performs well on linearly separable data
K-Nearest Neighbors	93%	Sensitive to feature scaling and K
Support Vector Machine (SVM)	90%	High margin separation; slower on large datasets
Decision Tree	91%	Tends to overfit without pruning
Random Forest	91%	Best performer, reduced variance

Confusion Matrix Format (for Logistic Regression)

[[65 3]

 [8 24]]

This means:

- 65 true negatives (didn’t purchase, correctly predicted)

- 24 true positives (purchased, correctly predicted)
- 3 false positives (predicted purchase, didn't purchase)
- 8 false negatives (missed purchase prediction)

Prediction Examples

Model predictions for various inputs (based on predict(sc.transform(...))):

Age	Salary	Logistic	KNN	SVM	Decision Tree	Random Forest
30	87,000	0	1	0	1	1
40	0	0	0	0	0	0
40	100,000	1	1	1	1	1
50	0	0	0	0	0	0
18	0	0	0	0	0	0
22	600,000	1	1	1	1	1
35	2,500,000	1	1	1	1	1
60	100,000,000	1	1	1	1	1

Graphical Analysis and Predictions

To visually interpret the classification models, **decision boundary plots** were generated using matplotlib. These plots illustrate how each model separates the classes (purchase vs. no purchase) in the feature space defined by **Age** and **Estimated Salary**.

- The **red region** represents customers predicted **not** to purchase insurance.
- The **green region** represents customers predicted **to** purchase insurance.
- Each data point is plotted with its actual class label.

Such visualizations were created for both **training** and **testing datasets**, allowing for inspection of model behavior and generalization.

Insights from the Graphs:

- **Logistic Regression**: Clear linear boundary; effective on well-separated data.
- **KNN**: Non-linear boundaries; slightly jagged edges due to neighbor sensitivity.
- **SVM**: Smooth linear boundary; shows strong class separation.
- **Decision Tree**: Sharp, step-wise boundaries; prone to overfitting.

- **Random Forest**: Smooth, general decision zones; best balance of flexibility and generalization.

Prediction Interpretation Table

Hypotheses and Assumptions

Based on preliminary data analysis and model outputs, the following hypotheses were formulated:

Hypothesis 1:

Younger individuals with higher salaries are more likely to purchase insurance.

Test:

- Prediction for Age 22, Salary ₹600,000 → Result: 1 (purchased)
- Prediction for Age 18, Salary not provided → Result: 0 (likely not purchased)

Supports the hypothesis — high salary can offset young age and increase purchase likelihood.

Hypothesis 2:

Older individuals with no salary are less likely to purchase insurance.

Test:

- Prediction for Age 50, No Salary → Result: 0
- Prediction for Age 60, Salary ₹100,000,000 → Result: 1

Supports the hypothesis — income plays a more critical role than age alone.

Hypothesis 3:

Salary has a stronger influence on insurance purchase behavior than age.

Test:

- Prediction for Age 35, Salary ₹2,500,000 → Result: 1
- Prediction for Age 40, No Salary → Result: 0

Confirmed — even modest age increase with no salary decreases likelihood, while high salary across age groups leads to positive predictions.

These hypotheses were validated using all five trained models, with **Random Forest** and **SVM** consistently aligning with these behavioral trends. This validates the model's reliability in capturing realistic patterns in user behavior.

Discussion

The comparative analysis of five classification algorithms — Logistic Regression, KNN, SVM, Decision Tree, and Random Forest — revealed several key findings regarding their effectiveness in predicting customer insurance purchases.

Model Performance Insights:

- **Random Forest** emerged as the best-performing model in terms of accuracy and generalization. It successfully minimized overfitting by averaging multiple decision trees, providing robust predictions across various customer profiles.
- **Logistic Regression** delivered strong performance for linearly separable data, validating its use as a reliable baseline model.
- **Support Vector Machine (SVM)** demonstrated consistent accuracy and was effective at clearly separating the classes, particularly after feature scaling.
- **KNN** performed well but showed signs of instability based on the value of K and was sensitive to scale and noise in the data.
- **Decision Tree**, while easy to interpret, showed signs of overfitting due to its depth and sensitivity to small data fluctuations.

Unexpected Observations:

- Some models predicted insurance purchases for very high salaries across all age groups, regardless of whether the user was very young or old, suggesting a **strong salary bias**.
- Inputs with **missing salary values** produced errors or poor predictions, highlighting the need for proper imputation or preprocessing in production environments.

Limitations:

- The dataset was limited to just **two features** (Age and Estimated Salary), which restricts model complexity and real-world applicability.
- No feature engineering or external data enrichment was performed, which could have improved prediction accuracy further.

Real-World Implications:

- These models can help insurance companies automate **lead scoring**, target marketing campaigns, and reduce manual effort in identifying potential buyers.
- The findings show the importance of **income level** over age in customer behavior modeling for insurance products.

Conclusion

This project focused on building and evaluating machine learning models to predict customer insurance purchases based on demographic features such as age and estimated salary. A total of

five classification algorithms were implemented and compared: **Logistic Regression**, **K-Nearest Neighbors**, **Support Vector Machine**, **Decision Tree**, and **Random Forest**.

Among these, **Random Forest** delivered the highest accuracy and robustness, making it the most suitable model for this use case. The project demonstrated that **estimated salary** has a more significant influence on purchase decisions than age, and that models can effectively identify customers most likely to purchase insurance.

The comparative approach used in this study offers valuable insights into the trade-offs between model complexity, performance, and interpretability. By analyzing real predictions and graphical boundaries, the project not only achieved its technical goals but also contributed actionable business intelligence that can help insurance companies improve targeting strategies.

Future enhancements may include:

- Incorporating more features such as education level, marital status, or region.
- Performing feature engineering or dimensionality reduction.
- Integrating models into a web app or dashboard for business users.

This project highlights the potential of AI to revolutionize decision-making in the insurance industry by enabling data-driven customer insights and prediction.

References

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
3. Zhang, Y., & Zhou, Z.-H. (2018). *A Review on Multi-Label Learning Algorithms*. IEEE Transactions on Knowledge and Data Engineering.
4. Brownlee, J. (2020). *Machine Learning Algorithms: From Scratch With Python*. Machine Learning Mastery.
5. Kaggle. (n.d.). *Social Network Ads Dataset*. https://www.kaggle.com/
6. Scikit-learn documentation. (n.d.). https://scikit-learn.org/stable/

Appendices

Appendix A: Code Repository

All models (Logistic Regression, KNN, SVM, Decision Tree, Random Forest) are organized in separate folders. Each folder contains:

- Model training code

- Predictions for given age-salary inputs
- Graphical decision boundary plots
- Accuracy and confusion matrix calculations

GitHub Link of the Project :

https://github.com/mVanisree/Customer-Insurance-Purchase-Prediction